AD_____

Award Number: DAMD17-98-1-8499

TITLE: The Single Prostate Cell Transcriptome as Biological Assay

PRINCIPAL INVESTIGATOR: Peter S. Nelson, M.D.

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

REPORT DATE: March 2002

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2002	Final (1 Sep 98 - 31 Jan 02)

4. TITLE AND SUBTITLE
The Single Prostate Cell Transcriptome as Biological Assay

5. FUNDING NUMBERS
DAMD17-98-1-8499

6. AUTHOR(S)
Peter S. Nelson, M.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

E-Mail: pnelson@fhcrc.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Technological advances in robotics and image analysis software have expanded the capabilities of large-scale strategies to provide for a more complete analysis of molecular differences between tissues or cells of interest. The research accomplished to date has demonstrated the ability to reproducibly isolate defined prostate cell populations by microdissection and flow cytometry. Gene expression studies of the cells purified by flow-cytometry reveal an altered expression profile that we believe results from the tissue dissociation/dispersion procedures. Ongoing and future work employs microdissection as the procedure of choice for specific cell-type analyses. The microdissection approach using a laser capture microscope is an efficient procedure for isolating cells representing abundant cell types, and we have isolated, purified, and analyzed the gene expression profiles from luminal epithelium and stromal elements. We have greatly expanded the database of sequences acquired from specific prostate cell types, and constructed arrays encompassing a wide range of diverse genes (n=6,000). In preliminary experiments we have used amplified cDNA probes isolated from small cell numbers to assess the gene expression profiles of defined cell types.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Prostate Cancer, transcriptome, cDNA, microdissection	38
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION

The subject of the research in this proposal is to develop methods for the examination of molecular alterations in prostate cancer at the level of homogeneous cell populations and single cells. The purpose of the research is to use these approaches to identify molecular alterations in prostate cancer cells that can be used either singly or in combination to provide insights into the molecular evolution of prostate carcinogenesis, and produce a set of molecular tools capable of influencing the clinical management of patients with prostate carcinoma. The scope of the research involved the construction of cDNA libraries representing the genes expressed in selected populations of normal and neoplastic prostate cancer cells followed by the construction of microarrays suitable for comprehensive gene expression studies. These arrays are then be used to evaluate methods for single-cell transcriptome amplification with the aim of identifying a cohort of cellular transcripts which reflect a cellular phenotype.

BODY

Technical objective 1: To *obtain defined populations of normal and neoplastic prostate cell types which retain in-situ cellular characteristics*

- *Task 1: obtain and pathologically characterize fresh samples of normal, primary neoplastic, and metastatic carcinoma. Prepare tissue sections in frozen and fixed formats. Perform immunohistochemistry.* Completed. A total of 20 samples were characterized as substrates for experiments in this project.

- *Task 2: purify normal luminal, normal basal, and primary carcinoma cell populations using flow cytometric sorting. Disaggregate tissues, immuno-label, sort, assess sorted populations for purity via microscopic examination and by PCR analysis. Sort single cells into microtiter format.* We have sorted and purified normal basal and luminal cells by flow cytometry and constructed a cDNA library from each population (described in the attached manuscript: Liu et al (2002)). We have sorted primary carcinoma cell populations as well (manuscript in preparation). Isolation of RNA from the purified cell populations has been inconsistent in terms of quality and quantity. We have optimized the methods using the RNA preservation agent RNAlater (Ambion Co).

- *Task 3: evaluate alternative tissue digestion protocols.* We have disaggregated tissue samples with trypsin, with EDTA alone, and with Dispase without a significant improvement in quality/quantity of RNA extraction compared to the standard collagenase protocol. Gene expression alterations resulting from the dis-aggregation procedure remain a major hurdle for using this approach with flow cytometry as a means to profile gene expression from solid tissues.

- *Task 4: microdissect cohorts of phenotypically distinct prostate cells: luminal epithelium, basal epithelium, PIN, carcinoma foci, metastatic foci.* We have employed a new approach for microdissection that uses a laser-capture microscope (Arcturus) and used this methodology to construct 3 prostate cDNA libraries; one representing prostate basal cells; one representing prostate luminal cells; and one representing prostate stromal cells. Following the de-

velopment of protocols aimed at optimizing both laser capture microdissection and RNA isolation, 24,000 cells each of stroma, luminal epithelium, and basal epithelium were captured and the RNA isolated by spin-column purification methods. cDNA libraries were constructed in a λ-phage vector using Clontech's SMART cDNA-PCR method. The respective libraries were then converted into phagemids and 300-700 clones from each library were sequenced for initial library characterization. Genes specific to each cell type were identified including PSA from the luminal cell library, PSCA from the basal cell library, and vimentin from the stromal cell library.

We have spent considerable time and effort optimizing the LCM approach for isolating pure cell populations in a form suitable for gene expression studies. Two major issues were discovered to be important when capturing cells from frozen prostate sections. One issue is the level of adhesion of the tissue section to the slide. The tissue section must be fixed and adhered onto the microscope slide such that it remains on the slide while performing hematoxylin staining and subsequent dehydrations, but not so adherent to the slide that cells will not be dissected from the slide when pulsed with the laser. In order to meet these demands, we have developed a simple method that consistently allows for efficient capture of cells from frozen prostate sections. Frozen prostate tissue embedded in OCT is first sectioned at 5 –10 μm in thickness onto clean, untreated microscope slides. Immediately following sectioning, slides are placed into 95% ethanol, without any exposure to air outside of the cryostat. Exposure of the slide to outside air will cause the section to adhere too tightly to the microscope slide for efficient dissection.

A second issue important for efficient laser-capture microdissection is the quality of the tissue section itself. If the section is wrinkled, has areas that come upward off of the slide, or even if the section is too small, the desired cells may not be efficiently captured. To deal with this problem, we use prep strips with a very light adhesive (Arcturus) to smooth out the tissue section. Sections that are both larger than the cap and prepared in this way present a smooth, even surface for the cap to set on. Uneven surfaces make focusing the laser onto the sections difficult and impede efficient capture.

Additional considerations must be made when the goal is to isolate RNA from laser-captured cells. A constant concern when isolating RNA is to eliminate RNases from preparations. To minimize RNase degradation of RNA, tissue is snap frozen in a liquid nitrogen/isopentane bath and stored at –80° C until cryosectioning at –20° C. When possible, a sample of the frozen tissue is checked for RNA quality prior to OCT embedding. To minimize the number of freeze-thaw cycles of a particular tissue block, the necessary number of sections are cut at one time from a single block, fixed in 95% ethanol made with DEPC-treated water, and stored in 100% ethanol at 4° C for three days while capturing is completed. Just prior to capturing cells from a section, the slide is rehydrated in DEPC-treated water, stained with a RNase-free solution of hematoxylin, rinsed again in DEPC-treated water, then dehydrated in two 100% ethanol baths, followed by two xylene baths. All baths are prepared fresh and all solution containers are treated to remove any contaminating RNases. Before placing a prepared slide that has been smoothed using adhesive preparation strips under the microscope, the microscope area is cleansed with an RNase inactivation solution (RNase-Zap, Ambion).

- *Task 5: microdissect single cells (20) from each of the above-described phenotypes.*

As described previously, while we have been able to consistently *isolate* single cells from prostate cancer sections using laser capture microdissection, the ability to amplify the amount of cDNA needed for use in cDNA library construction or cDNA microarray analysis from the limited amount of RNA available in single cells remains challenging. We are continuing to investigate techniques, including amplification of mRNA that will allow us to develop comparative gene expression profiles from individual prostate cells. At the same time, we have made progress in developing the tools necessary for the analyses of molecular changes within individual prostate epithelial cells isolated from peripheral blood, apheresis samples, and/or bone marrow. These tools include immunostaining cell preparations (epithelial cells isolated from peripheral blood, apheresis samples, or bone marrow using positive and negative selection methods) for prostate specific antigen, capturing positively-stained cells by laser capture microdissection, cell lysis, and DNA analysis for methylation of GST-pi and androgen receptor mutations. We have successfully amplified and sequenced exon 8 of the androgen receptor from individual circulating prostate cells and have identified no mutations in a cohort of 5 patients. We are currently developing a protocol to determine GST-pi methylation status.

- *Task 6: assess RNA quality (preservation) between frozen sections and fixed/stained sections.* As anticipated, our work in this area has demonstrated that the yield of RNA from frozen tissues is much greater and of higher quality than from comparable quantities of formalin fixed tissue. Our current protocol employs a rapid ethanol fixation of frozen tissue with or without an H&E or immunostain prior to LCM. We have successfully isolated intact RNA from formalin-fixed tissues, but to date this remains poorly reproducible.

- *Task 7: assess feasibility of flow sorted single cell isolation automation.* We are not currently pursuing this approach due to the alterations in gene expression resulting from tissue disaggregation. Future work may entail flow cytometric isolation of epithelial cells in peripheral blood or bone marrow.

- *Task 8: Refine cell phenotype acquisition based upon the development of new markers/antibodies.* In collaboration with Dr. Alvin Liu in the Department of Molecular Biotechnology, we have identified several additional antigens recognized with monoclonal antibodies that can be used for sorting prostate epithelial cells by flow cytometry (see reportable outcomes, Liu et al). The future application of these discriminating proteins/antigens will await the development of consistent amplification protocols as described in this proposal.

Technical objective 2: *To construct microarrays of prostate transcripts that reflect the gene expression potential of the cell types to be examined.*

- *Task 8: identify a non-redundant clone set from the Prostate Expression Database to encompass all highly expressed transcripts (~12), moderately expressed transcripts (~500) and several thousand rare transcripts (~6000).*
  We have now identified and assembled a non-redundant set of 6,000 cDNAs (ESTs) from the prostate expression database that are suitable for array construction. Many of these genes are derived from the cell type-specific libraries described above.

- *Task 9: retrieve cDNA clones from archive, PCR amplify inserts with amine-linked primer, and purify.* We have retrieved 6,000 cDNA clones from the cDNA archive and amplified the inserts by PCR. Our current array construction methodology at the Fred Hutchinson Cancer Center uses poly-lysine-coated slides and demonstrates excellent reproducibility and sensitivity. We have used these arrays for the identification of genes in the prostate under the control of the androgen receptor and androgenic ligands (See reportable outcomes, Nelson et al).

- *Task 10: construct 3 normalized cDNA libraries from flow sorted basal, luminal, and primary carcinoma (CD44+) without amplification procedures, and evaluate libraries for quality: diversity and abundance of transcripts.*
As described in the previous report, we have constructed cDNA libraries from flow sorted basal (CD44+), luminal (CD57+), and primary carcinoma (CD44+) cells. We have also now constructed cDNA libraries from normal basal and luminal epithelial cells using microdissection approaches. A total of 2,500 ESTs have been produced from these libraries and entered into the Prostate Expression Database (www.pedb.org).

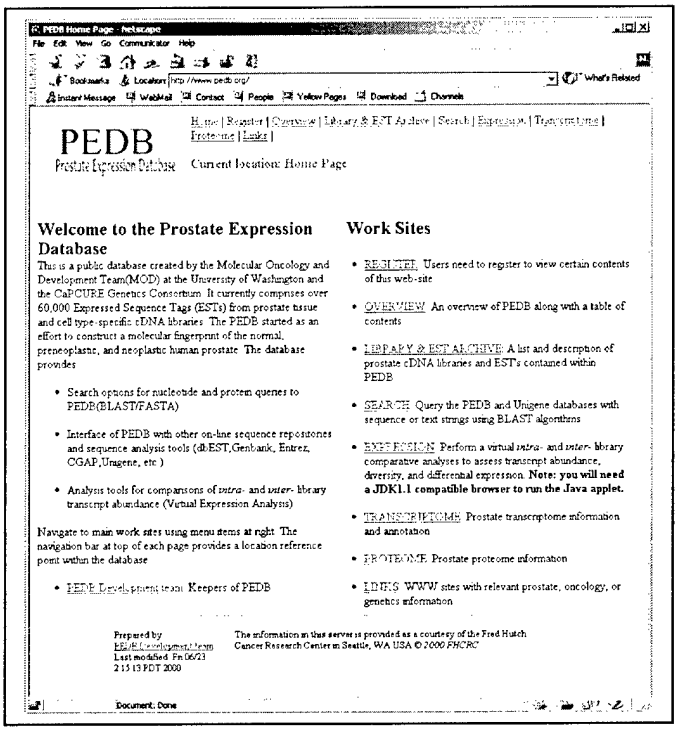


Figure 1. PEDB Web Interface. ESTs derived from cDNA clones sequenced from the microdissected prostate cell types are submitted to the Prostate Expression Database (PEDB) for processing. The sequences are assembled and annotated. A non-redundant set of 6,000 ESTs representing 6,000 different genes expressed in prostate tissues was identified. The cDNA clones were redistributed in microtiter plates, amplified, and spotted onto glass microscope slides for microarray hybridization experiments.

- *Task 11: pick random cDNA clones from the new libraries, array on nylon membranes and screen for abundant prostate cDNAs, select non-abundant species, PCR amplify inserts.* Random sequencing of cDNA clones from the libraries described above and additional libraries from prostate cancer cell lines has identified >18,000 distinct genes expressed in prostate tissues. We have used a virtual selection approach rather than the physical negative selection approach, to identify non-redundant clone sets representing the prostate transcriptome. These clones have been extracted from the database archive, re-arrayed into 384-well microtiter plates, amplified by PCR, and spotted onto microscope slides for subsequent hybridization.

- *Task 12: construct physical micro-arrays of cDNA clones on glass supports using robotic tools: total of 500 replicates.* See Task 9 above. We are currently using a GeneMachines robotic spotting tool with the capability of spotting >18,000 cDNAs per microscope slide. More than 500 replicate slides have been printed to date comprising the 6,000 prostate PEDB cDNAs. The current use for these slides is for the analysis of amplification procedures in order to assess the fidelity of probe material obtained from small numbers of cells.

- *Task 13: assess alternative array methodologies as they become available (ink jet oligonucleotide)* To date, in our hands, the spotted cDNA microarray approach has the attributes of the greatest versatility (ability to rapidly customize when new cDNAs are identified), low cost, and accuracy. Future work will involve the continued analysis of alternative array platforms.

Technical objective 3: *To construct representative probes from single or small numbers of defined cells that are suitable for micro-array interrogation, and retain the transcriptome composition (diversity and abundance) present in the original cell type(s).*

- Task 14: convert to cDNA, amplify by PCR, and label nucleic acid from flow sorted cell populations of decreasing cell quantities. Assess quality by Northern analysis and hybridization to small "known clone" array. Compare with unamplified "traditional" probe. (months 12-13). We are not currently using the flow-cytometry isolation approach due to changes in gene expression associated with tissue disruption. Our studies have utilized microdissection.
- Task 15: as above with microdissected populations. (months 13-14). We have successfully microdissected prostate luminal and basal cells from 10 μm frozen sections. Amplification using the PCR-based strategy incorporating an anchored primer has been successful in producing adequate amounts of cDNA for probe construction and hybridization. However, the fidelity of the amplification in numerous attempts has been poor. This results in the skewing of message abundance levels in the probe material relative to the starting material. One attempt has been made to modify this approach using truncated cDNAs followed by adapter ligation and subsequent amplification. This approach did not produce better results than the whole cDNA amplification. However, we have not abandoned this idea, and modifications are planned in the future.
- Task 16: as above with aRNA method and flow sorted cells (months 15-16). As described above, we are not currently using flow-sorting for cell isolation and probe construction.

- Task 17: as above with microdissected populations. (months 17-18). We have used a modification of the aRNA protocol developed by Eberwine et al. We have achieved a ~1000-fold amplification with a first round aRNA synthesis and an additional ~1000-fold amplification with a second round. We have evaluated this approach using 'in-house' reagents as well as commercially available kits from Arcturus and Ambion. To date the Ambion methods have produced the most consistent results. This allows for the use of ~0.5 ng of total RNA for probe construction. We have utilized ~20,000 cells in analyses of gene expression levels by microarray. However, in our hands, the aRNA amplification is still not suitable for the analysis of single, or small numbers (<10,000) of microdissected cells.

Gene expression analyses from LCM cells. Following RNA isolation from LCM prostate cells, several methods to analyze gene expression were evaluated. These methods include cDNA library construction/analysis, three-prime end amplification, and linear RNA amplification.

cDNA library construction. In order to identify novel genes expressed in the human prostate, as well as to contribute to the understanding of the gene expression profiles within the prostate, we have developed three cDNA libraries from cells of the normal human prostate captured using laser capture microdissection technologies. Following the development of protocols aimed at optimizing both laser capture microdissection and RNA isolation, stroma, luminal epithelium, and basal epithelium (24,000 cells each) were captured and the RNA isolated by spin-column purification methods. cDNA libraries were constructed in a λ-phage vector using Clontech's SMART cDNA-PCR method. The respective libraries were then converted into phagemids and up to 768 clones from each library were sequenced for initial library characterization (Table 1).

Table 1. Normal human prostate stroma, luminal epithelium, and basal epithelium were isolated by laser capture microdissection, RNA was isolated and cDNA constructed using the SMART cDNA PCR library construction kit (Clontech). cDNA was cloned into the λTriplEx vector and isolated clones were sequenced by the dideoxy sequencing chain termination method. Numbers in parenthesis represent the number of annotated clones.

	Luminal Epithelium	Basal Epithelium	Stroma
Number of clones sequenced	768	288	741
% without annotations	27	55	28
% annotated	73 (557)	45 (130)	72 (534)
% mitochondrial	18	34	11
% ribosomal	3	10	3.7

To date, there are no published reports characterizing or comparing and contrasting the gene expression profiles and/or cDNA library construction from pure populations of prostate stroma, luminal epithelial, or basal epithelial cells. We expect that these libraries will be useful tools for a variety of applications, including identifying prostate-specific genes, cell-type specific genes within the prostate, and in differential gene expression analysis. To further streamline this method, a modified protocol was developed which eliminated phage cloning. Instead, PCR-amplified cDNA library inserts were directly cloned into plasmid. Initial experiments performed with this protocol demonstrated more efficient PCR amplification and sequencing of clones and indicate that this method will be useful in the generation of future libraries.

RNA amplification. An inherent problem encountered with our cDNA libraries was the necessity, again due to the limiting amounts of RNA available following LCM, to have an amplification step. The SMART-PCR libraries had up to 36 PCR amplification cycles included in the generation of cDNA clones. This was an impediment to identifying unique, or rare, clones. In an attempt to overcome this issue, the problem was tackled by experimenting with protocols designed initially by Eberwine, *et al*, to linearly amplify the RNA, rather than logarithmically amplifying the cDNA. Briefly, this technique begins with limiting amounts of starting total RNA (e.g. less than 2 μg) and, following reverse transcription using a T7-oligo d (T) primer, second strand cDNA synthesis and ds cDNA purification, and finally RNA generation using *in vitro* transcription, results in the production of up to 10^3 fold amplification of RNA as antisense RNA (aRNA). More aRNA can be generated by additional rounds of amplification, yielding 10^6 fold amplification of starting RNA. We, and others, have found that for this technique to be useful it is essential that the starting RNA be of very high quality. In addition, we and others have observed the necessity for the incorporated T7-oligo d(T) primer to be HPLC purified by particular companies to ensure the necessary high quality. These issues were found by us to be just the tip of the iceberg in regards to the technical details required regarding the linear amplification of RNA from LCM cells.

We have had success in amplifying RNA from as few as 2,000 LCM cells. Over 70 μg of aRNA was generated after only two rounds of RNA amplification. This aRNA, while relatively short (averaging 200-600 bp in length), has proven to be useful in quantitative RT-PCR and cDNA microarray hybridizations. We have found that successful RT-PCR requires the use of primers annealing very near to the 3' end of the aRNA. Using optimized primers, we were able to detect not only genes known to be highly expressed in the human prostate (GAPDH, prostate specific antigen [PSA]), but also genes expressed at low levels in the prostate (retinal short-chain dehydrogenase 3, bone morphogenetic protein 6, DOPA decarboxylase, and dihydrofolate reductase). The expression of PSA was found to be, as expected, more highly expressed in luminal epithelium than stroma (up to 2000 fold more) and basal epithelium (25 fold). Successive rounds of RNA amplification do appear to decrease the ability to detect gene expression, presumably largely due to the short length of the resulting cDNA generated using random hexamers and aRNA which has been amplified in 2 or 3 successive rounds. We have found, however, that the greatest decrease in ability to detect gene expression occurs between rounds one and two, rather than between two and three rounds of amplification. While a limited number of groups have shown that the representative RNA population after one or two rounds of RNA amplification remains reflective of the unamplified mRNA population, we believe this requires additional experimentation. Currently, we are continuing extensive validation of this protocol using human prostate RNA.

We have been able to reproducibly demonstrate that RNA from LCM cells can yield up to 80 μg of aRNA after two or three rounds of RNA amplification and that the resulting aRNA can be used in quantitative RT-PCR, as described above, and in cDNA microarray hybridizations. Using cDNA microarray hybridization, we have found that the correlation coefficient between the log ratio of two different populations of cells from either amplified or unamplified starting RNA is approximately 0.7, while the correlation coefficient between one and two rounds of RNA amplification from identical starting RNA was over 0.9. This is reasonably close to the inherent variability and reproducibility when using microarrays to analyze gene expression. For example, we have found that the coefficient correlation between replicate hybridizations (starting with the same RNA preparation, but being separately reverse transcribed and amplified) to be between 0.90 and 0.94; the correlation coefficient between duplicate samples (starting with separate microdissections and RNA isolations) was found to be 0.81. Often, the correlation coefficient of log ratios between two different cell populations can be less than 0.3.

To date, there are no published reports characterizing or comparing and contrasting the gene expression profiles and/or cDNA library construction from populations of prostate stromal, luminal epithelial, or basal epithelial cells. A manuscript detailing these libraries is in preparation (see reportable outcomes, Moore et al). We anticipate that these libraries will be useful tools for a variety of applications, including identifying prostate-specific genes, cell type-specific genes within the prostate, and in differential gene expression analysis. Library construction from PIN, primary carcinoma, and metastatic carcinoma are in progress.

- Task 18: as above with microdissected populations from frozen and fixed tissues. (months 19-20). We have extensively used LCM for acquiring pure cell populations from frozen tissues. Our efforts in extracting RNA from archived formalin tissues have not been successful. We are currently evaluating the optimal length of formalin fixation anticipating that short fixation times may be compatible with RNA preservation.
- Task 19: convert to cDNA, amplify, label, and hybridize single-cell probes to high-density oligonucleotide arrays. (months 21-25). As described above, we have been unable to achieve reproducible amplification from the quantities of RNA contained in a single cell.
- Task 20: capture and quantitate hybridization spot intensities on fluorimage laser scanners, and enter into database. (months 21-25). We have acquired microarray database software compatible with the PEDB Oracle platform and have enabled a preliminary system for archiving microarray data. We are in the process of linking the microarray data with PEDB EST data and the annotated genomic sequence data available from the UCSC genome assembly.

Technical objective 4: *To identify a cohort of cellular transcripts which correlate with, define, or "fingerprint", a cellular phenotype(s).*

- Task 21: examine hybridization intensities (values) for each datapoint in an automated, comparative fashion from cells of *a priori* defined identical phenotype (luminal epithelium with luminal epithelium) to develop cohorts of phenotype-defining transcripts.

- Task 22: examine hybridization intensities between cells with *a priori* defined different phenotypes to establish a lineage relationship. (months 26-27). Microarray hybridizations were performed with probes generated from microdissected populations of luminal epithelium, basal epithelium and stromal cells. (Figure 2).We were able to generate reproducible hybridization fingerprints for each cell phenotype.

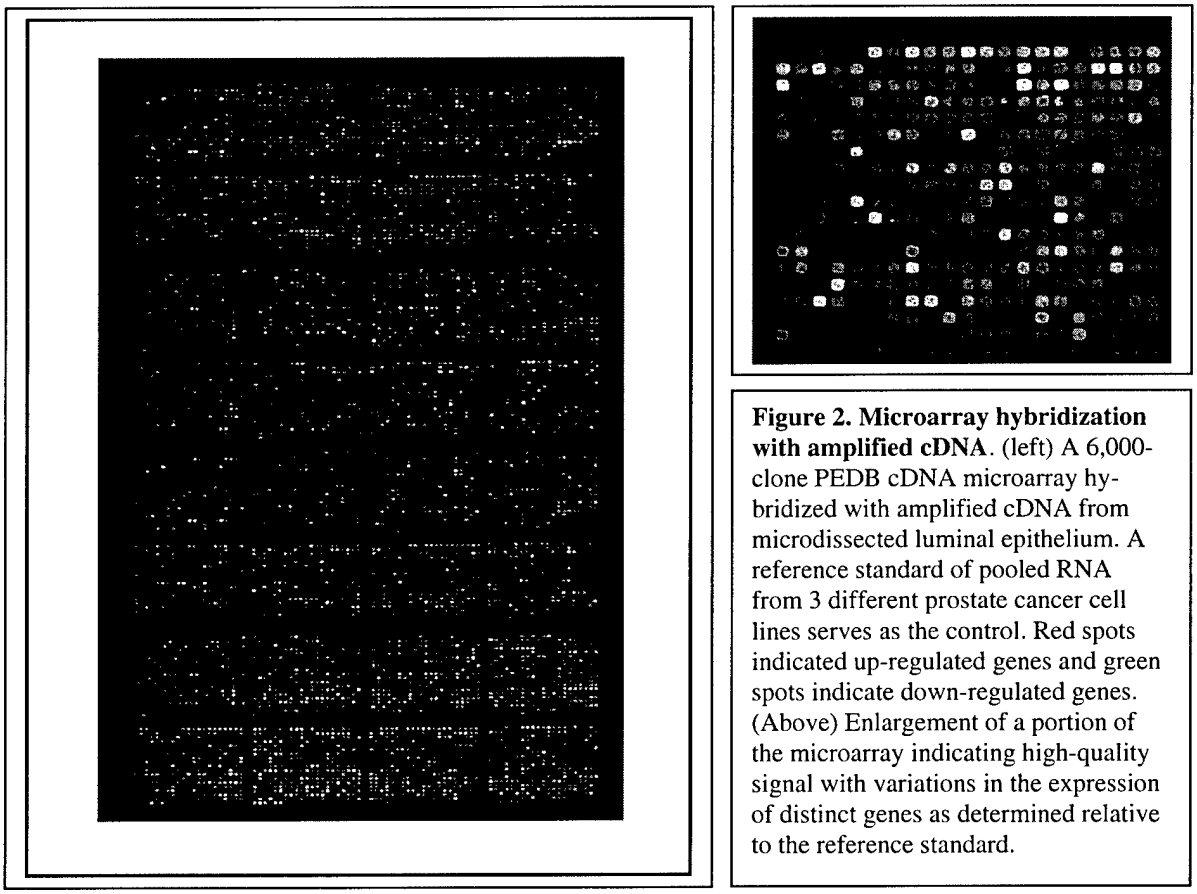


Figure 2. Microarray hybridization with amplified cDNA. (left) A 6,000-clone PEDB cDNA microarray hybridized with amplified cDNA from microdissected luminal epithelium. A reference standard of pooled RNA from 3 different prostate cancer cell lines serves as the control. Red spots indicated up-regulated genes and green spots indicate down-regulated genes. (Above) Enlargement of a portion of the microarray indicating high-quality signal with variations in the expression of distinct genes as determined relative to the reference standard.

- Task 23: correlate expression profiles with known molecular/biochemical/functional data concerning each cell type. (months 27-28). Table 2 provides a preliminary assessment of genes that are differentially-expressed in each of the 3 phenotypically-defined cell populations characterized by microarray expression: luminal epithelium, basal epithelium, and stromal cells. Several of these genes have previously been characterized as cell type-specific (e.g. PSA, PAP: luminal epithelium). Other cDNAs represent known genes that have not been associated with a particular cell type or represent uncharacterized transcripts.
- Task 24: analyze by DNA sequencing cDNAs which are in phenotype cohorts and have not previously been defined. (months 26-28). Ongoing work involves the full-length cDNA cloning of cDNAs that consistently cluster with one specific cell type.
- Task 25: analyze expression data using cluster and phylogeny algorithms to assess lineage relationships. (months 26-29). Ongoing work involves the acquisition of replicate expression profiles of microdissected cell populations in order to determine consistent profiles that would enhance the confidence placed in exclusively assigning gene expression to a particular cell type/compartment. To date we have analyzed 2 prostate specimens (luminal cell, basal

cell, and stroma). We have ongoing work to analyze an additional 4 samples prior to cluster analysis and phylogenetic determination.

- Task 26: plan molecular experiments and clinical evaluation of candidate phenotype-defining cohorts: e.g.1) retrospective analysis of carcinomas with known clinical outcomes (progression/metastasis) 2) prospective analysis diagnostic needle biopsy samples 3) evaluation of unrecognized or "latent" cancer samples obtained at autopsy. (months 27-30). We have started to acquire 'fingerprints' of prostate carcinomas and to date we have determined the gene expression profiles of 6 including 4 prostate cancer xenografts. Based on the heterogeneity we have seen, we estimate that we will need to acquire a total of 20 samples in order to have a statistical basis to categorize samples as: 'basal cell like-' or 'luminal cell like-' according to their expression profiles. At that juncture, a clinical correlation with outcomes will be assessed.

Table 2: Identification of differentially expressed cDNAs in specific prostate cell phenotypes using cDNA microarray analysis. Microdissected cell populations were used as templates for probe construction. Each cDNA was represented by 4 datapoints to ensure reproducibility.

Luminal epithelium	Basal epithelium	Stroma
PSA	Sorbitol dehydrogenase	Myosin light chain kinase
PAP	Gastrin-binding protein	Skeletal muscle LIM-protein FHL1
B_2 microglobulin	ACLP-aortic carboxypeptidase	KIAA0353
Complement factor B	Bone marrow stromal cell antigen 2	EST264
Znalpha2GP	Desmin	Alpha-actin
LNCaP0578	Prostaglandin D synthase	Matrix Gla protein
Cytokeratin 18	EST262	Thymosin beta 10
EST 195	SM22	Prostacyclin-stimulatory factor
EST213	Insulin-like growth factor binding protein	Alpha-tropomyosin
Prostase	Alpha-2-glycoprotein-1	Alpha-2 collagen type IV
Glutathione-insulin transhydrogenase	LNCaP0257	Insulin-like growth factor-1
SNC19	EST-similar to CREB-RP	Basic calponin
h-lamp-2	g2053071	Caldesmon
Calgizzarin	ESTs57	Basic fibroblast growth factor
KIAA0438 (neurodegeneration protein)	LNCaP2043	hevin

- Task 26: analyze/compile data and prepare formal report (month 30).

Our initial gene expression analyses in prostate luminal epithelium, basal epithelium, and stroma indicates that this will be a powerful method to identify not only genes unique to a particular cell type, but also to define a gene expression profile which may lead to a more comprehensive understanding of the cellular function of various cell populations and the cellular pathways active within them. Further analyses of prostate basal epithelium will be

particularly interesting as this cell type appeared, by initial observations, to resemble that of a myoepithelial cell type; an observation controversial in the current literature. Technical hurdles have limited our ability to assess single-cell populations for their intrinsic expression profiles. However, our future objectives will incorporate technological advances with the microarray resources that we have developed in the context of this proposal and we anticipate applying these tools toward the phenotyping of individual prostate cancer cells.

KEY RESEARCH ACCOMPLISHMENTS

- Obtained and purified single circulating neoplastic prostate cells from the peripheral blood of patients with prostate cancer and analyzed exon 8 of the androgen receptor for molecular alterations.
- Constructed cDNA libraries from laser-capture microdissected prostate luminal and basal epithelial cells and prostate stroma.
- Sequenced and analyzed 2,000 cDNAs (producing ESTs) from the luminal cell, basal cell, and stromal libraries.
- Performed virtual analyses to identify genes differentially expressed in these distinct cell types.
- Constructed cDNA microarrays comprised of 6,000 different prostate-derived cDNAs.
- Acquired and implemented database software for archiving and analyzing cDNA microarray experiments.
- Constructed complex cDNA probes from microdissected cells and used these probes in microarray hybridization experiments.
- Identified cDNAs (known and novel) with differential expression in distinct cell type of luminal, basal, and stromal elements.

REPORTABLE OUTCOMES

Abstracts:

Biaoyang L, Ferguson C, White J, Hood L, and **Nelson PS** (1999) Identification of genes regulated by androgens in human prostate carcinoma cell line LNCaP. MTAP User Meeting. Monterey, CA. (abstract)

Nelson PS, Hawkins V, Ferguson C, Vessella R, Lange P, and Hood L. (1999) The Prostate Expression Database: A Tool for Novel Gene Discovery and the Virtual Analysis of Prostate Gene Expression. Proceedings of the American Urological Association. (abstract)

Manuscripts:

Nelson PS, Han D, Rochon Y, Corthals G, Lin B, Monson A, Nguyen V, Franza BR, Plymate SR, Aebersold R, and Hood L. (2000) Comprehensive analyses of prostate gene expression: convergence of EST databases, transcript profiling and proteomics. *Electrophoresis* 21:1823-31.

Grouse LH, Munson PJ, and **Nelson PS**. (2001) Sequence Databases and Microarrays as Tools for Identifying Prostate Cancer Biomarkers. *J. Urology* 57 (Suppl 4A): 154-159.

Li P and **Nelson PS**. (2001) Prostate Cancer Genomics. *Current Urology*.

Clegg N, Erolgu B, Ferguson C, Arnold H, Mooreman A, and **Nelson PS**. (2002) Digital Expression Profiles of the Prostate Androgen Response Program. *J Steroid Bioch and Mol Biol.* 80: 13-23.

Liu AY, **Nelson PS**, van den Engh G, Hood L. (2002) Human Prostate Epithelial Cell-Type cDNA Libraries and Expression Pattern in Prostate Cancer. *The Prostate.* 50:92-103. (note that this work was supported by the present proposal though not explicitly stated in the acknowledgements).

Nelson PS, Clegg N, Arnold H, Ferguson C, Bonham M, White J, Hood L, and Lin B. The Program of Androgen-Responsive Genes in Neoplastic Prostate Epithelium (Submitted).

Moore S and **Nelson PS.** Comparative Analysis of the Luminal and Basal Epithelial Cell Transcriptomes from the Normal and Neoplastic Prostate. (In preparation).

CONCLUSIONS

The research accomplished in the context of this project has demonstrated the ability to reproducibly isolate defined prostate cell populations by microdissection and flow cytometry. Gene expression studies of the cells purified by flow-cytometry reveal an altered expression profile that we believe results from the tissue dissociation/dispersion procedures. Until more appropriate dispersion procedures are developed for solid tissues (in contrast to cells in body fluids), our procedure of choice for isolating defined cell types is microdissection. The microdissection approach using a laser capture microscope is an efficient procedure for isolating cells representing abundant cell types, and we have isolated, purified, and analyzed the gene expression profiles from luminal epithelium and stromal elements. To date, the microdissection approach is not efficient in the isolation of sparse/rare cell populations. We have greatly expanded the database of sequences acquired from specific prostate cell types, and constructed arrays encompassing a wide range of diverse genes (n=6,000). We have used amplified cDNA probes, isolated from small cell numbers acquired by LCM, to assess the gene expression profiles of defined cell types. One objective of our proposal that was not achieved centered on determining the gene expression profile of *single* cells. We feel that this is a very important objective, and will continue to pursue these studies. However, using larger cell numbers isolated by LCM, we have identified several candidate genes that are differentially expressed between luminal epithelium and basal epithelium and between epithelial and stromal cell types of the prostate. The confirmation of these results is ongoing and has provided a foundation for additional studies aimed at identifying the role(s) of these cell type-specific genes in normal and neoplastic prostate development.

REFERENCES

None

APPENDICES

Clegg N, Erolgu B, Ferguson C, Arnold H, Mooreman A, and **Nelson PS**. (2002) Digital Expression Profiles of the Prostate Androgen Response Program. *J Steroid Bioch and Mol Biol.* 80: 13-23.

Liu AY, **Nelson PS**, van den Engh G, Hood L. (2002) Human Prostate Epithelial Cell-Type cDNA Libraries and Expression Pattern in Prostate Cancer. *The Prostate.* 50:92-103. (note that this work was supported by the present proposal though not explicitly stated in the acknowledgements).

PERGAMON

Journal of Steroid Biochemistry & Molecular Biology 80 (2002) 13–23

The Journal of Steroid Biochemistry & Molecular Biology

www.elsevier.com/locate/jsbmb

Digital expression profiles of the prostate androgen-response program

Nigel Clegg, Burak Eroglu, Camari Ferguson, Hugh Arnold, Alec Moorman, Peter S. Nelson*

Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA

Received 15 June 2001; accepted 24 September 2001

Abstract

The androgen receptor (AR) and cognate ligands regulate vital aspects of prostate cellular growth and function including proliferation, differentiation, apoptosis, lipid metabolism, and secretory action. In addition, the AR pathway also influences pathological processes of the prostate such as benign prostatic hypertrophy and prostate carcinogenesis. The pivotal role of androgens and the AR in prostate biology prompted this study with the objective of identifying molecular mediators of androgen action. Our approach was designed to compare transcriptomes of the LNCaP prostate cancer cell line under conditions of androgen depletion and androgen stimulation by generating and comparing collections of expressed sequence tags (ESTs). A total of 4400 ESTs were produced from LNCaP cDNA libraries and these ESTs assembled into 2486 distinct transcripts. Rigorous statistical analysis of the expression profiles indicated that 17 genes exhibited a high probability ($P > 0.9$) of androgen-regulated expression. Northern analysis confirmed that the expression of *KLK3/PSA*, *FKBP5*, *KRT18*, *DKFZP564K247*, *DDX15*, and *HSP90* is regulated by androgen exposure. Of these, only *KLK3/PSA* is known to be androgen-regulated while the other genes represent new members of the androgen-response program in prostate epithelium. LNCaP gene expression profiles defined by two independent experiments using the serial analysis of gene expression (SAGE) method were compared with the EST profiles. Distinctly different expression patterns were produced from each dataset. These results are indicative of the sensitivity of the methods to experimental conditions and demonstrate the power and the statistical limitations of digital expression analyses.

Keywords: Androgen; Prostate; EST; SAGE; Transcriptome

1. Introduction

Genes regulated by androgenic hormones are of critical importance for the normal physiological function of the human prostate gland, and they contribute to the development of prostate diseases such as benign prostatic hypertrophy (BPH) and prostate carcinoma. Androgens such as testosterone and dihydrotestosterone (DHT) interact with the androgen receptor (AR) leading to the transcriptional activation of androgen-target genes [1]. This gene network regulates prostate morphogenesis, growth, and function, and promotes the development and progression of prostate neoplasia [2]. Despite the importance of androgens in modulating diverse prostate cellular processes, relatively few components of this androgen-response program have been identified or characterized.

Current estimates indicate that between 35,000 and 40,000 genes are encoded in the human genome [3,4]. To confer developmental and functional specificity, only a fraction of this total is transcribed in a given tissue or cell type at any given time. This repertoire of expressed genes in transcript form is termed the transcriptome [5], a dynamic assessment or inventory of gene expression activity that reflects the cellular developmental state and response(s) to environmental perturbations. Proceeding from the hypothesis that comprehensive gene expression profiles will provide insights into cellular function, several procedures have been developed to qualitatively and quantitatively assess transcriptomes. These methods can be broadly divided into analog approaches

Abbreviations: *KLK3*, kallikrein 3; *RPLP0*, ribosomal protein large, P0; *UQCRC2*, ubiquinol-cytochrome *c* reductase core protein 2; *FKBP5*, FK506-binding protein 5; *DKFZP564K247*, DKFZP564K247 protein; *PHGDH*, phosphoglycerate dehydrogenase; *KRT18*, keratin 18; *RPS25*, ribosomal protein S25; *EIF3S6*, eukaryotic translation initiation factor 3, subunit 6 (48 kDa); *FTL*, ferritin, light polypeptide; *DDX15*, DEAD/H (Asp-Glu-Ala/His) box polypeptide; *RPS27A*, ribosomal protein S27A; *ACADVL*, acyl-coenzyme A dehydrogenase, very long chain; *KIAA0101*, KIA0101 gene product; *DKFZP564D0462*, hypothetical protein DKFZP-564D0462; *RPS15A*, ribosomal protein S15a; *DED*, apoptosis antagonizing transcription factor; *BSG*, basigin; *TPI1*, triosephosphate isomerase 1; *CLTB*, clathrin, light polypeptide (Lcb); *DBI*, diazepam binding inhibitor; *ENO1*, enolase 1 (alpha); *KLK2*, kallikrein 2; *KLK4*, kallikrein 4; *ODC1*, ornithine decarboxylase 1; *PDHA1*, pyruvate dehydrogenase (lipoamide) alpha 1; *TMEPA1*, transmembrane, prostate androgen-induced RNA; *TUBA1*, tubulin, alpha 1; *UGT2B17*, UDP glycosyltransferase 2 family, polypeptide B17; *VEGF*, vascular endothelial growth factor

* Corresponding author. Fax: +1-206-667-2917.

E-mail address: pnelson@fhcrc.org (P.S. Nelson).

0960-0760/02/$ – see front matter
PII: S0960-0760(01)00167-4

such as DNA array analysis [6–8], and digital methods as exemplified by expressed sequence tag (EST) quantitation [9] and the serial analysis of gene expression (SAGE) [10]. Each approach has distinct advantages and limitations that have been detailed previously [11]. A principle advantage of digital methods is the possibility of sampling the complete transcriptome in a single experiment. These approaches also permit the analysis of previously uncharacterized genes and allow for direct statistical analyses of transcript numbers rather than relying on indirect measures of transcript ratios.

Our objective in this study was to identify genes expressed in human prostate cells exhibiting transcriptional regulation by androgens. We hypothesize that such genes could be direct mediators of the androgen-receptor pathway or be involved in prostate-specific functions that could be exploited for understanding normal and neoplastic prostate growth. To facilitate systematic studies of prostate gene expression, we have established the prostate expression database (PEDB), an archive that contains more than 70,000 ESTs generated from prostate cDNA libraries [12]. Two libraries constructed specifically for this study comprise genes expressed in the LNCaP prostate cancer cell line under conditions of androgen stimulation and androgen deprivation. The LNCaP cell line represents a model system for the study of androgen regulation as LNCaP cells express a functional AR, proliferate in response to physiological levels of androgens, and increase the transcription of known androgen-regulated genes such as prostate specific antigen (PSA) [13]. We applied statistical tools to compare these EST datasets and identified both known and novel genes with a high probability ($P > 0.9$) of being regulated by androgens. Northern analysis was used to confirm androgen-regulated expression. These studies identified *FKBP5*, *KRT18*, *DKFZP564K247*, *DDX15*, and *HSP90*, as new members of the prostate epithelial androgen-response program. LNCaP transcriptomes defined by two distinct SAGE experiments were also examined for genes exhibiting androgen regulation and these results were compared with the EST profiles. These results support the use of comprehensive gene expression profiling methods to define cellular responses to hormonal stimuli, and demonstrate both the power and the statistical limitations of digital expression analyses.

2. Materials and methods

2.1. Cell culture

The prostate carcinoma cell line LNCaP was obtained from ATCC and grown in RPMI 1640 with 10% FCS (Life Technologies, Inc.). Cells were transferred into RPMI-1640 medium with 10% charcoal-stripped fetal calf serum (CS-FCS) 24 h before androgen-regulation experiments. This medium was replaced with fresh CS-FCS media or fresh CS-FCS including 1 nM of the synthetic androgen R1881 (New England Nuclear Life Science Products, Inc.). Cells were harvested for RNA isolation at 0- and 24-h time points.

2.2. Library construction

Total RNA was isolated from androgen-stimulated (LNCaP01) and androgen-starved (LNCaP02) cells using TRIzol (Life Technologies, Inc.) according to the manufacturer's instructions. Poly(A)$^+$ RNA was purified using oligo(dT) chromatography [14]. A unidirectional library was constructed in the pSport1 vector (Life Technologies, Inc.) according to a modification of the Gubler and Hoffman [15] protocol. Poly(A$^+$) was reverse-transcribed using superscript reverse transcriptase and an oligo(dT) linker/primer containing a *Not*1 site (Life Technologies). Sephacryl-S400 (Pharmacia) was used to size-select the synthesized cDNA and remove excess linkers. Blunt-ended, double-stranded cDNA was ligated with a *Sal*1 adapter, digested with *Not*1, then ligated into *Sal*1–*Not*1 digested pSport1. High-efficiency electrocompetent *Escherichia coli* were transformed using a Bio-Rad GenePulser under recommended conditions. Approximately, 86% of the LNCaP01 and 89% of the LNCaP02 transformants contained inserts. The average insert size for the library was 1.7 kb.

2.3. DNA sequencing

Independent transformant colonies were picked into 100 ul PCR mix [10 mM Tris, pH 8.3, 1.5 mM $MgCl_2$, 50 mM KCl, 120 uM dNTPs, 1 U Taq polymerase (Promega) and 0.12 uM each of VN26 TTTCCCAGTCACGACGTTGTA and VN27 GTGAGCGGATAACAATTTCAC] and subjected to 40 cycles of 30 s at 94 °C, 30 s at 60 °C and 120 s at 72 °C followed by 10 min at 72° C. Amplified inserts were purified over Sephacryl S-500 (Pharmacia), and 4 ul was used in DNA sequencing reactions using M13 reverse fluorescent-labeled dye primers as detailed in the Prism cycle sequencing kit (Applied Bio-systems, Inc.). Reaction products were electrophoresed on ABI 373 and 377 DNA sequencers.

2.4. Northern analysis

Total RNA was isolated from LNCaP cells using the TRIzol method according to the manufacturer's instructions. Ten micrograms of total RNA was fractionated on 1.2% agarose gels under denaturing conditions and transferred to nylon membrane using the capillary method. Blots were hybridized with cDNA probes labeled with [α-^{32}P]-dCTP using a Random Primers DNA labeling kit (Life Technologies Inc.) according to the manufacturer's protocol. Filters were imaged and quantitated using a phosphor-capture screen and Image Quant software (Molecular Dynamics). β-Actin was used as an internal control for normalizing transcript levels between samples.

2.5. EST assembly, annotation, and comparison

DNA sequences were stored, clustered, and annotated using the PEDB relational database management tools and data analysis pipeline [17]. [1] Briefly, vector, *E. coli*, and interspersed repeats were masked in the ESTs using Cross_Match [2] and Repeatmasker. [3] Poor quality sequences, with >50% ambiguous nucleotides ('N') between nucleotides 100 and 500 were discarded. CAP2 [16], a multiple sequence alignment program based on a variant of the Smith–Waterman algorithm, was used to cluster the masked sequence and generate a consensus sequence for each assembly. Each distinct cluster was annotated by searching Unigene, [4] GenBank, [5] and dbEST [6] databases using BLASTN. [7] Annotations were assigned automatically using SmartBlast (Perl 5.0) to select the database match with the lowest *P*-value and the highest BLAST score where the maximum *P*-value was e^{-20} and the minimum BLAST score was 500. Some species required manual reconciliation when either two distinct PEDB species were annotated with the same identification, or when annotations differed between public databases. The Virtual Expression Analysis Tool (VEAT [8]) and scripts written in Perl 5.0 were used for creating transcript species reports. The biological role for each species was assigned using the categories described by Adams et al. [9]. Supplemental information, including a complete list of species and transcript frequencies is available at the PEDB web site. Gene symbols are from the HUGO Gene Nomenclature Committee.

Using statistics described by Audic and Claverie [11], differential gene expression in androgen-stimulated and androgen-deprived cells was inferred based on differential representation of ESTs in cDNA libraries.

2.6. SAGE data acquisition and analysis

The following LNCaP SAGE libraries are listed at the NCBI Library Browser web site [9] and were downloaded from SAGE-map's anonymous FTP site [10]: SAGE_Chen_LNCaP (62,681 tags), SAGE_Chen_LNCaP_no-DHT (65,206 tags), SAGE_CPDR_LNCaP-C (41,848 tags), and SAGE_CPDR_LNCaP-T (44,370 tags). For simplicity, these libraries are hereafter called LNCaP(+)DHT, LNCaP(−)DHT, LNCaP-C and LNCaP-T. Statistical analyses were performed using the software provided at the SAGEmap xProfiler web site. [11]

[1] http://www.pedb.org.
[2] http://www.genome.washington.edu/UWGC/methods.htm.
[3] http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker.
[4] ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.seq.all.Z.
[5] ftp://ncbi.nlm.nih.gov/blast/db/nt.Z.
[6] ftp://ncbi.nlm.nih.gov/blast/db/est.Z.
[7] http://blast.wustl.edu.
[8] http://www.pedb.org.
[9] http://www.ncbi.nlm.nih.gov/SAGE/sagelb.cgi.
[10] ftp://ncbi.nlm.nih.gov/pub/sage/seq/.
[11] http://www.ncbi.nlm.nih.gov/SAGE/sageexpsetup.cgi.

3. Results

3.1. EST-derived LNCaP transcriptomes

Two cDNA libraries, LNCaP01 and LNCaP02, were constructed from the prostate adenocarcinoma cell line LNCaP under conditions of androgen stimulation and androgen starvation, respectively. Approximately, 2300 ESTs were produced from each library and the sequences were entered into the PEDB [12]. Automated processing of the ESTs to remove short, poor quality, repetitive, and/or vector sequences eliminated 779 ESTs from further analysis. The remaining 4458 ESTs were assembled using the CAP2 sequence assembly program. Each EST cluster was annotated by searching the Unigene, GenBank, and dbEST databases with the CAP2-generated cluster consensus sequences using BLASTN. Clusters annotated with the same database sequence were joined, and all ESTs grouped to the same cluster were assigned the same unique PEDB cluster ID. ESTs for mitochondrial genes were grouped as a single cluster and accounted for approximately 6% of all ESTs. These genes were not further analyzed. In total, 2486 distinct transcript species were identified (Fig. 1): 2240 were homologous to previously identified genes or ESTs, and 252 were not significantly homologous to any public database sequence. The latter species may represent novel genes or previously unsequenced regions of known genes.

The number of distinct transcripts comprising the LNCaP01 and LNCaP02 transcriptomes are quantitatively similar, but qualitatively different. In all, 87% of the species were represented in one transcriptome or the other, but not in both (Fig. 1A). Despite the difference in species composition, the EST frequency distributions of the two samples were similar: nearly 78% of the species are represented by a single EST and only 9% were composed of more than 2 ESTs (Table 1). These distributions are broadly consistent with previous estimates which indicate there are relatively few transcripts expressed in high abundance (5–15 species at 10,000 copies per cell), an intermediate number of moderately abundant transcripts (500 species at 300 copies per cell) and many low abundance transcripts (10,000 different species expressed in 1–15 copies per cell) [17]. In all, 70% of the transcript species with two or more ESTs in either LNCaP01 or LNCaP02 were also present in the other library (Fig. 1B). Thus, while few low abundance transcripts were found in both datasets, most of the high abundance transcripts were found in common.

Functional roles were assigned to each distinct species according to the convention established by Adams et al. [9]. The five primary biological roles were cell division, cell signaling/cell communication, cell structure/motility, cell/organism defense, and metabolism. For graphical presentation, we added the 'androgen-regulated' category to emphasize the primary difference between the experimental samples (Fig. 2). In total, 923 transcript species could be

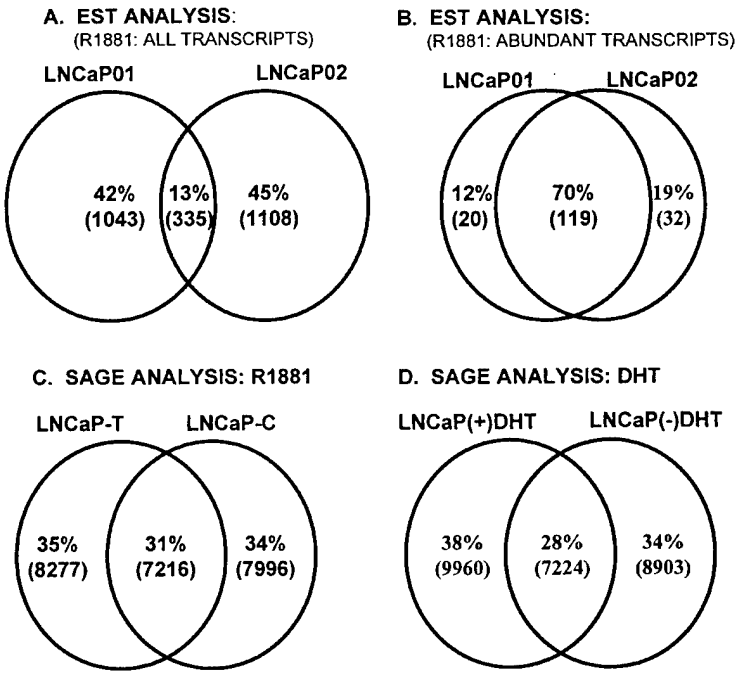


Fig. 1. Summary of LNCaP transcriptome diversity determined by EST and SAGE analysis. Representations of (A) the EST-derived number of all distinct transcripts unique to two LNCaP cell states (synthetic androgen R1881-stimulated LNCaP, LNCaP01; and R1881-starved LNCaP, LNCaP02) and those expressed in common between the two cell states; (B) the EST-derived number of highly and moderately expressed transcripts in LNCaP01 and LNCaP02 (>2 ESTs in one or both libraries) and those expressed in common; (C) SAGE analysis determining the number of distinct transcripts unique and in common between R1881-stimulated and starved LNCaP cells; (D) SAGE analysis determining the number of distinct transcripts unique and in common between DHT-stimulated and starved LNCaP cells.

Table 1
Distribution of molecular species by EST frequency

ESTs/species	No. of species (proportion of total)	
	LNCaP01	LNCaP02
1	1064 (0.78)	1133 (0.79)
2	202 (0.15)	199 (0.14)
3	55 (0.04)	56 (0.04)
4	26 (0.02)	23 (0.02)
5	8 (0.01)	8 (0.01)
6	6 (<0.01)	8 (0.01)
>6	17 (0.01)	16 (0.01)
Total	1378	1443

assigned biological roles. A detailed annotation of LNCaP transcripts assigned to these functional roles can be viewed at the PEDB website. [12] Both LNCaP transcript profiles have a similar distribution of species in each functional category (Fig. 2). The protein/gene expression category is the largest, primarily because of the high frequency of ESTs for ribosomal proteins and translation factors. Similar results have been obtained for whole normal prostate tissue [18]. A comparison of the composition of broad functional categories does not reveal a cohort of genes that reflect androgen stimulation or starvation, but differential gene expression in response to androgens is clearly evident for individual genes (Fig. 2). *KLK3/PSA*, an androgen-regulated gene, represents 1.4% of the ESTs in LNCaP01 (derived from androgen-stimulated cells), but only 0.05% of the ESTs in LNCaP02. ESTs for the androgen-response genes *KLK2*, *KLK4*, *ODC1*, *TUBA1*, and *ENO1* were also more abundant in the LNCaP01 library.

3.2. Androgen-regulated genes identified by digital expression analysis

We compared the abundance of each transcript species represented in the androgen-stimulated and androgen-starved transcriptomes using a VEAT [12]. VEAT provides a comprehensive graphical view of transcript frequency, as defined by EST number, between two or more transcriptomes of interest (Fig. 3). Among the species with more than two ESTs in either library, the most extreme difference in EST frequency was observed for *KLK3/PSA*. Twenty-nine *KLK3/PSA* ESTs were isolated from LNCaP01, the library made from androgen-stimulated cells, and only one EST was isolated from LNCaP02 (Table 2). This finding was expected as *KLK3/PSA* is one of the most abundant transcripts

[12] www.pedb.org.

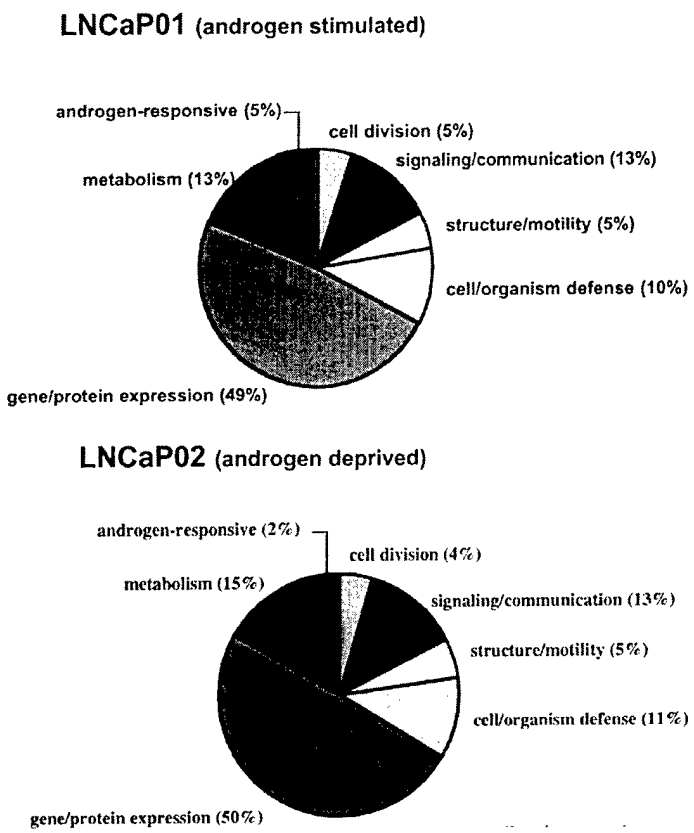


Fig. 2. Functional categorization of the LNCaP cell transcriptome. EST assemblies were annotated against the Genbank and Unigene databases. A putative functional role was assigned based upon categories developed by TIGR (http:www.tigr.org) and the percentage of ESTs corresponding to each role are depicted under cellular conditions of androgen stimulation and androgen starvation.

in the prostate [18] and is known to be transcriptionally regulated by androgens in LNCaP cells.

Additional differences in EST frequencies were seen for many other LNCaP transcripts. Determining the significance of these observations is challenging because of the potential for chance events (e.g. randomly selecting a given cDNA clone from a library) when the event is part of a large population of observable outcomes (e.g. cDNA libraries are complex and comprised of millions of cDNA clones). In order to validate and prioritize more subtle differences in gene expression, we used a statistical approach designed to provide a confidence interval indicating the probability that a given set of observations could occur by chance, or alternatively represents a significant change in expression [11]. Software available on the Internet [13] computes the confidence intervals corresponding to arbitrary significance levels and sample sizes of two datasets N_1 and N_2 [11]. Twenty-one species were predicted to be differentially expressed with a probability exceeding 90%: 9 were increased in response to androgens, and 12 were increased by androgen starvation (Table 2). With the exception of *KLK3/PSA*, none of these genes has previously been reported to be androgen-regulated in the prostate.

To confirm the differential expression statistics, the levels of transcription of *KLK3/PSA* and nine additional genes were examined by Northern analysis (Table 2, Fig. 4). cDNAs representing five different transcripts predicted to be androgen-upregulated by EST analysis were hybridized to Northern blots of RNA extracted from androgen-starved and androgen-stimulated LNCaP cells. Transcipts from each of the five genes were more abundant in androgen-stimulated cells than in androgen-deprived cells. Consistent with the EST frequency data, *KLK3/PSA* expression was increased 35-fold in androgen-stimulated cells compared to androgen-starved cells (Fig. 4). The transcripts encoding keratin 18 (*KRT18*), a gene expressed in prostate secretory cells, were increased 5-fold. FK506 binding protein 5 (*FKBP5*), *DKFZP564K247*, and *UOCRC2* were induced to a lesser extent. In contrast, statistical predictions were inaccurate for four of five putatively down-regulated genes. The steady-state level of DKFZp564K247 RNA was actually increased by androgens, and reduced transcription of

[13] http://igs-server.cnrs-mrs.fr.

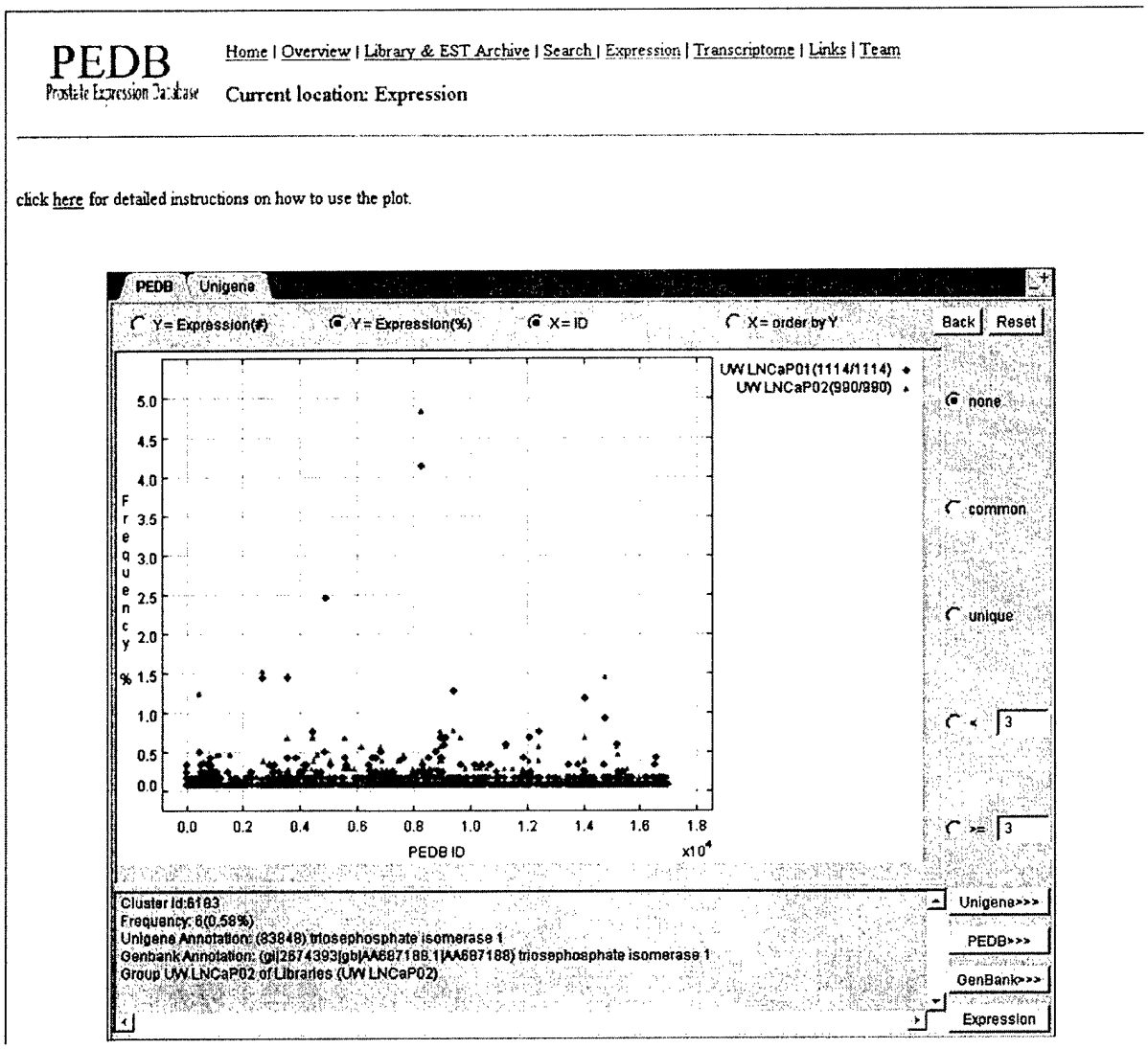


Fig. 3. Virtual differential expression determined by digital expression profiles. A view of cellular gene expression using the VEAT from the PEDB. Distinct transcripts are assigned a unique database ID and ordered along the *X*-axis. The number of ESTs assembled into each unique transcript (frequency) is displayed on the *Y*-axis as a percentage of the total EST number obtained from each library. Each library is represented by a different symbol (e.g. LNCaP01, triangle; LNCaP02, diamond). Highlighting any data point (using a mouse) provides annotation corresponding to that particular transcript (PEDB reference).

eukaryotic initiation factor 3 subunit 6 (*EIF3S6*), ribosomal protein 27a (*RPS27A*), and basigin (*BSG*) was not confirmed by Northern analysis. Surprisingly, one gene predicted to be decreased by androgen deprivation, the RNA helicase DEAD/H box polypeptide 15 (*DDX15*), was upregulated more than 3-fold by Northern analysis. There are several RNA helicases and our probe may be cross-hybridizing with another closely related androgen-inducible gene. At least, one other androgen-regulated RNA helicase has been reported [19].

In addition to the six androgen-responsive genes identified above, a heat shock protein gene (*HSP90*) was initially identified as androgen-regulated after a preliminary statistical analysis of approximately 1500 LNCaP01 and LNCaP02 ESTs. As the number of ESTs increased, *HSP90* was not differentially expressed based on the arbitrary statistical probability cut-off of $P > 0.90$; however, Northern blot analysis demonstrated a 4-fold increase in *HSP90* expression with androgen stimulation. There are numerous genes in the heat shock protein 90 gene family with strong sequence similarity [20], and our Northern hybridization conditions cannot differentiate between them. Nevertheless, this result confirms that one or more members of the *HSP90* gene family are androgen-responsive.

Table 2
Putative androgen regulated genes in LNCaP01/LNCaP02 libraries ($P \geq 0.9$) and corresponding SAGE data

Gene	ESTs			Androgen Response on Northern blot[a]	SAGE		
	No. of ESTs		Probability of differential expression[b]		SAGE Tag[c]	Probability of differential expression[d,e]	
	LNCaP01[f]	LNCaP02[g]				LNCaP-T/-C[h]	LNCaP(+)DHT/(−)DHT[i]
KLK3/PSA	29	1	$P > 0.99$	+35	GGATGGGGAT	$P = 1.00$ (82/5)	$P = 0.25$ (63/36)
RPLP0	22	9	$0.98 < P < 0.99$	nd[j]	CTCAACATCT	$P = 0.00$ (120/105)	$P = 0.00$ (248/292)
UQCRC2	5	0	$0.96 < P < 0.97$	+1.3	AAAGTCAGAA	$P = 0.16$ (6/8)	$P = 0.16$ (6/5)
FKBP5	4	0	$0.93 < P < 0.94$	+1.9	GTTCCAGTGA	$P = 0.66$ (6/0)	$P = 0.39$ (0/2)
DKFZP564K247	4	0	$0.93 < P < 0.94$	+1.7	TATCGGGAAT	–	$P = 0.29$ (2/1)
PHGDH	4	0	$0.93 < P < 0.94$	nd	TTACCTCCTT	$P = 0.22$ (22/12)	$P = 0.15$ (65/40)
KRT18	4	0	$0.93 < P < 0.94$	+5.0	CAAACCATCC	$P = 0.12$ (22/14)	$P = 0.02$ (27/35)
RPS25	6	1	$0.93 < P < 0.94$	nd	AATAGGTCCA	$P = 0.00$ (53/51)	$P = 0.06$ (132/84)
SFTPD	9	3	$0.90 < P < 0.91$	nd	–	–	–
EIF3S6	0	6	$0.98 < P < 0.99$	+1.2	AATATTGAGA	$P = 0.07$ (11/10)	$P = 0.33$ (12/6)
FTL	0	5	$0.96 < P < 0.97$	nd	CCCTGGGTTC	$P = 0.24$ (9/15)	$P = 0.15$ (22/37)
DDX15	0	4	$0.93 < P < 0.94$	+3.5	ATCGTTGTAA	$P = 0.37$ (4/1)	$P = 0.47$ (3/0)
RPS27A	0	4	$0.93 < P < 0.94$	+1.3	AACTAACAAA	$P = 0.15$ (16/10)	$P = 0.14$ (49/31)
ACADVL	0	4	$0.93 < P < 0.94$	nd	GCCGCCCTGC	$P = 0.13$ (6/6)	$P = 0.48$ (8/20)
KIAA0101	0	4	$0.93 < P < 0.94$	nd	ATGATTTATT	$P = 0.21$ (3/4)	$P = 0.47$ (3/0)
DKFZp564D0462	0	4	$0.93 < P < 0.94$	−2.6	CAGTTCTCAC	$P = 0.29$ (1/1)	$P = 0.40$ (2/0)
RPS15A	0	4	$0.93 < P < 0.94$	nd	GACAAAAAAA	$P = 0.26$ (27/14)	$P = 0.18$ (12/8)
RPS15A	–	–	–	–	GACTCTGGTG	$P = 0.16$ (11/7)	$P = 0.00$ (36/41)
DED	0	4	$0.93 < P < 0.94$	nd	GCACCTATTG	$P = 0.29$ (2/1)	$P = 0.35$ (0/1)
Species1145	0	4	$0.93 < P < 0.94$	nd	–	–	–
BSG	1	6	$0.92 < P < 0.93$	−1.02	GCCGGGTGGG	$P = 0.06$ (11/11)	$P = 0.00$ (216/341)
TPI1	1	6	$0.92 < P < 0.93$	nd	TGAGGGAATA	$P = 0.01$ (33/29)	$P = 0.02$ (39/32)

[a] Ratio of normalized signal intensity from RNA of hormone stimulated/starved cells.
[b] [11].
[c] Most abundant unique tag.
[d] [35].
[e] Tag frequency in hormone stimulated/starved samples.
[f] 2222 ESTs.
[g] 2236 ESTs.
[h] ~42,000 tags per library.
[i] ~62, 000 tags per library.
[j] nd, not done.

3.3. Comparison of EST and SAGE digital expression profiles

An alternate method of acquiring qualitative and quantitative transcript profiles is by the SAGE. Rather than producing gene tags of 300–500 nucleotides, the SAGE method generates sequence tags of approximately 10 nucleotides in length. This difference allows 10–30-fold more SAGE tags to be acquired per sequencing reaction, thus, deeper transcript profiles can be obtained more efficiently. However, the short tag length may introduce ambiguity when assigning a tag to a specific gene [21].

Data from two independent SAGE profiling experiments examining androgen-regulated gene expression in LNCaP cells were obtained from the SAGEmap website at NCBI.[14] Descriptions of the libraries indicated that one SAGE dataset, designated LNCaP(−)DHT/(+)DHT, was derived from LNCaP cells grown in hormone-depleted media for 3 months (LNCaP(−)DHT) and then stimulated with 1 nM DHT (LNCaP(+)DHT) for 24 h. Approximately 63,000 tags were sequenced from each library. The second SAGE dataset, LNCaP-T/-C, was derived from cells grown in hormone-depleted media for 5 days (LNCaP-C), then stimulated with 10^{-8} M R1881 for 24 h (LNCaP-T). Approximately, 42,000 tags were sequenced from each library. The distribution of expressed genes in each pair of SAGE libraries is given in Fig. 1B and C.

Theoretical and empirical data suggest that roughly 650,000 transcripts must be sampled to identify all but very rare mRNAs in the cell [22]. Thus, neither our study nor the SAGE datasets were large enough to thoroughly sample transcript diversity in the LNCaP cells, and neither dataset is capable of identifying differential gene expression among low abundance transcripts. Broadly, genes with a role in protein synthesis (ribosomal proteins and translation initiation factors) were the most abundant transcripts in both our EST data and the SAGE profiles. Interestingly, the EST approach identified approximately 200 transcript species

[14] http://www.ncbi.nlm.nih.gov/SAGE/.

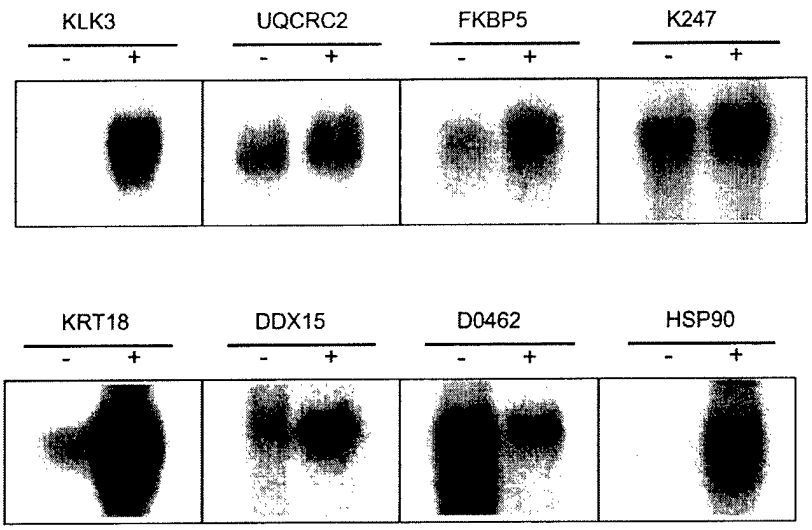


Fig. 4. Northern blots of eight androgen regulated genes predicted to be differentially expressed by virtual EST analysis. K247 is *DKFZP564K247* and D0462 is *DKFZP564D0462*. 'Minus', total RNA from androgen-starved LNCaP cells. 'Plus', total RNA from LNCaP cells treated with 1 nM R1881.

with corresponding Unigene entries that were not observed in the SAGE libraries. Conversely, the SAGE studies identified hundreds of transcripts that were not observed in the EST assemblies. Thus, these studies complement each other in creating an inventory representing the LNCaP cell transcriptome.

Transcripts with a high probability of differential expression between each pair of SAGE profiles were identified using the SAGEmap xProfiler. Despite a 10-fold difference in sample size, the SAGE and EST studies identified similar numbers of putative androgen-responsive genes (cut-off $P = 0.9$). In the EST analysis, 21 genes had a high probability

Table 3
Known androgen-response genes exhibiting differential expression in one or more libraries ($P \geq 0.6$)

Gene	ESTs			SAGE tag[a]	SAGE		
	No. of ESTs		Probability of differential expression[b]		Probability of differential expression[c,d]		Prostate-enriched[e]
	LNCaP01[f]	LNCaP02[g]			LNCaP-T/-C[h]	LNCaP(+)DHT/(−)DHT[i]	
CLTB	0	0	$0.00 < P < 0.10$	GGCTGGGCCT	$P = 0.45$ (3/0)	$P = 0.73$ (2112)	−
DBI	1	0	$0.50 < P < 0.60$	TGTTTATCCT	$P = 0.77$ (13/2)	$P = 0.03$ (20/18)	−
ENO1	6	3	$0.60 < P < 0.70$	GTGTCTCATC	$P = 0.13$ (9/12)	$P = 0.04$ (15/14)	−
KLK2	3	0	$0.80 < P < 0.90$	CTGTGGTTTA	$P = 0.39$ (2/0)	$P = 0.80$ (810)	+
	−	−	−	CTGTGGTTAA	−	$P = 0.76$ (14/3)	+
KLK3	29	1	$P > 0.99$	GGATGGGGAT	$P = 1.00$ (82/5)	$P = 0.25$ (63/36)	+
KLK4	2	0	$0.70 < P < 0.80$	AAATTGACCC	$P = 0.35$ (1/0)	$P = 0.51$ (2/8)	+
ODC1	4	1	$0.70 < P < 0.80$	TGCGTGGTCA	$P = 0.35$ (1/0)	−	−
	−	−	−	ATGCAGCCAT	−	$P = 0.11$ (7/7)	−
PDHA1	0	0	$0.00 < P < 0.10$	CAGTTTGTAC	$P = 0.60$ (5/0)	$P = 0.28$ (4/2)	−
PMEPA1[j]	1	0	$0.50 < P < 0.60$	TGATGTCTGG	$P = 1.00$ (29/1)	$P = 0.47$ (7/2)	+
TUBA1	4	1	$0.70 < P < 0.80$	GAGGAGGGTG	$P = 0.29$ (2/4)	$P = 0.44$ (5/13)	−
UGT2B17	0	0	$000 < P < 0.10$	GAGGGTTTTA	$P = 0.62$ (0/5)	$P = 0.40$ (4/1)	−
VEGF	1	1	$0.00 < P < 0.10$	TTTCCAATCT	$P = 0.29$ (1/2)	$P = 0.69$ (610)	−

[a] Most abundant unique tag.
[b] [11].
[c] [35].
[d] Tag frequency in hormone stimulated/starved cells.
[e] More abundant in the prostate than in most other tissues.
[f] 222 ESTs.
[g] 2236 ESTs.
[h] ~42,000 tags per library.
[i] ~62,000 tags per library.
[j] Tag inferred from [34].

of differential expression (9 up-regulated, 12 down-regulated) while 17 unique tags were identified in the SAGE LNCaP-T/-C study (6 up-regulated, 11 down-regulated), and 23 were identified in the SAGE LNCAP(+)DHT/(−)DHT study (17 up-regulated, 6 down-regulated). Surprisingly, with the exception of *KLK3/PSA*, all of the identified genes were different across the three datasets. *KLK3/PSA* had a high probability of differential expression in both our EST dataset ($P > 0.99$) and the LNCAP-T/-C dataset ($P = 1.0$). The only other potential androgen-regulated gene in the EST data that had a moderate probability of differential expression based on SAGE was FK506 binding protein 5 (*FKBP5*; $P = 0.66$, LNCaP-T/-C). The three genes that we confirmed to be differentially expressed by Northern blot analysis (keratin 18, 3-phosphoglycerate dehydrogenase, and *DKFZP564K247*) were not expressed at significantly different levels ($P < 0.30$) in the two SAGE datasets.

A review of published literature identified 75 genes reported to be androgen-responsive in one or more human tissues (see PEDB [15]). Twenty-three of these genes had corresponding EST tags; 47 had LNCaP-T/-C SAGE tags; and 55 had LNCaP(+)DHT/(−)DHT SAGE tags. Thus, SAGE sampling of 10-fold more transcripts only doubled the number of observed, previously-described, androgen-regulated genes. The genes identified in the EST dataset are not just a subset of those found in the larger SAGE datasets: *TMPRSS2*, a serine protease gene whose transcription is stimulated by androgen in LNCaP cells [23], was represented in the EST data, but not in the SAGE libraries. Only 12 of the 75 known androgen-response genes had even a moderate probability of differential expression ($P \geq 0.6$) in one or more datasets (Table 3), and there is no case where statistical predictions agree across all three data sets. Six of the twelve genes were predicted to be androgen inducible in the EST dataset, compared to five genes in the LNCaP-T/-C dataset and three in the LNCaP(+)DHT/(−)DHT dataset. The two SAGE studies, with similar numbers of tags, predicted completely different cohorts of up-regulated genes (Table 3).

4. Discussion

The identification and quantitation of the complement of genes expressed in a cell or tissue provides a framework for understanding biological properties and establishes a tool set for functional studies. Several methods have been developed for the comprehensive analysis of gene expression in complex biological systems. We have investigated the application of two procedures, EST profiling and SAGE, to characterize the transcriptome of prostate adenocarcinoma cells and to identify the cohort of genes regulated directly or indirectly by androgenic hormones. The EST profiles obtained from two LNCaP cDNA libraries identified 2486 distinct transcripts. Of these, 336 were expressed in common. The total number of transcripts, we identified in this study represents about 12–17% of the total complexity found in prostate epithelium [24] and likely includes all highly expressed, many moderately expressed and relatively few rarely expressed transcripts. Many of these genes were previously identified in other tissues, but were not known to be expressed in the prostate. In all, 252 new transcripts were identified that are not represented in any public database. Since over 2.2 million human ESTs are present in dbEST (release 081800), some of the unknown transcripts may be exclusively expressed in the prostate epithelium. These findings support the continued utility of cataloging transcripts from specialized tissue sources. These newly identified cDNAs can be tested for tissue-specific expression and can be used both to facilitate the identification of exons in the context of the human genome project and to enhance the positional cloning of prostate cancer susceptibility genes.

Androgens regulate numerous processes in prostate epithelial cells that include cell division, cell quiescence, apoptosis, lipid metabolism, and the production of specialized secretory proteins such as KLK3/PSA. Of the 2486 distinct transcripts identified in the LNCaP transcriptome, 364 (14%) showed at least a 2-fold difference in expression following exposure to androgens. Statistical analysis reduced this number to 21 genes with a high probability of differential expression ($P \geq 0.9$). Ten were further tested by Northern analysis which confirmed six were indeed transcriptionally regulated by androgen; *KLK3/PSA*, *FKBP5*, *KRT18*, *DDX15*, and *DKFZP564D0462*. In addition, *HSP90* was identified as an androgen-response gene by Northern blot analysis. These data identify five genes as new members of the androgen-response network, since only *KLK3/PSA* was previously known to be androgen-responsive. The lack of complete concordance between the digital expression results and Northern analysis can be partly explained by cross-hybridization to highly-homologous gene family members, alternative splicing events, and the lack of Northern sensitivity to alterations in low abundance transcripts.

The genes found in this study to be transcriptionally sensitive to androgen have diverse functions. *KLK3/PSA* is a highly abundant serine protease with known androgen-response elements in the promoter region [25] and prostate-enriched expression. Keratin 18 is a marker for prostate luminal cells [26] but is found in a variety of epithelia. The *DKFZP564D0462* gene encodes a putative seven transmembrane-domain protein that is expressed in a variety of tissues. The DEAD/H box polypeptide 15 gene is a putative RNA helicase similar to a yeast gene required for mRNA splicing [27]. Another RNA helicase, GRTH, is up-regulated in testis in response to androgen [19]. These genes may play a role in steroidogenesis or androgen-mediated stimulation of protein synthesis. *HSP90* binds and activates the androgen receptor. *FKBP5*, another gene predicted to be up-regulated in LNCaP cells, interacts with *HSP90* in func-

[15] http://www.pedb.org.

tionally mature progesterone complexes [28]. Hence, both *HSP90* and *FKBP5* may be up-regulated to facilitate signal transduction through the androgen receptor.

While general trends in gene expression were similar with respect to the overall effects of androgens, why was little concordance found between EST data and the SAGE data in terms of the expression of specific genes? In part this may be attributable to relatively small overall sample sizes and the limitations of statistical confidence. Cloning or sequencing biases could be unequally introduced by the experimental approaches, and ambiguity in SAGE tag assignment may affect a subset of genes. However, an alternative explanation is that each method accurately reflects the state of cellular gene expression, and the differences are attributable to the actual in vitro conditions. There will be some variation in transcript levels even under optimal conditions that may relate to cell density, growth media, and other factors. At present, we do not know the precise effects of protracted androgen starvation on LNCaP cells, but the extended starvation of cells used to create the LNCaP(+)DHT/(−)DHT libraries (3 months), could have selected for altered gene expression. In this regard, it is noteworthy that *KLK3/PSA*, one of the most abundant androgen regulated genes, was not differentially expressed in the LNCaP(+)DHT/(−)DHT dataset (Table 3). Cell-line history may also affect transcription. LNCaP may have undergone significant physiological adaptation and genomic change during maintenance in different laboratories. Esquenet et al. [29] observed a marked decrease in the ability of androgen to induce *KLK3/PSA* transcription in LNCaP cells of high passage number relative to cells of low passage number. And LNCaP cells can undergo "proliferative shut-off" in response to androgen [30]. These experimental differences may be analogous to the heterogeneity observed between individual cancers and may be reflected in the cellular transcriptomes assayed by digital-expression profiles.

Another intriguing possibility is that different androgens and androgen concentrations activate or repress subnetworks of the androgen-response program. Testosterone, DHT, and synthetic androgens such as R1881 induce a concentration-dependent biphasic growth response in LNCaP cells that may be influenced by the relative activities of growth-promoting and growth-suppressing genes [31]. Different ligands or ligand concentrations may recruit distinct AR co-activator molecules that dictate the subset of genes to be activated [32,33]. Of interest, a report describing the cloning and characterization of the gene corresponding to the SAGE tag exhibiting the greatest androgen-induction (29-fold) in the LNCaP(+)DHT/(−)DHT SAGE dataset was recently published [34]. By Northern analysis, the expression of this gene, *PMEPA1*, was shown to increase only 2-fold with 10^{-10} M R1881, but nearly 5-fold with 10^{-8} M R1881; the concentration used in the SAGE experiments. The 10^{-9} M R1881 concentration used in our EST experiments did not induce a detectable increase in *PMEPA1* EST frequency.

At present, financial and technological barriers make it impractical to simultaneously test all known genes for expression in the prostate. Inventories of genes from cell lines such as LNCaP, which are used extensively as model systems for studying prostate cancer, can help alleviate this problem by identifying the subset of genes of relevant to the biological system under study. Additional SAGE and EST data are needed to identify rare transcripts and to increase statistical power required for robust digital expression studies. In addition to their demonstrated utilities as gene discovery and analysis tools, the digital expression profiling methods used here can also greatly facilitate the construction of microarray-based reagents suitable for applications where higher throughput is required.

Acknowledgements

We thank Barbara Trask for critical review of the manuscript, David Haynor and Roger Bumgarner for helpful discussions, and Sue Heiner for manuscript preparation. This work is supported in part by grants (DAMD 17-98-1-8499 and PC991274) from the Department of Defense, a Grant (CA75173) from the National Cancer Institute, and the CaPCURE foundation.

References

[1] A.O. Brinkmann, L.J. Blok, P.E. de Ruiter, P. Doesburg, K. Steketee, C.A. Berrevoets, J. Trapman, Mechanisms of androgen receptor activation and function, J. Steroid Biochem. Mol. Biol. 69 (1999) 307–313.

[2] J. Trapman, K.B. Cleutjens, Androgen-regulated gene expression in prostate cancer, Seminars Cancer Biol. 8 (1997) 29–36.

[3] B. Ewing, P. Green, Analysis of expressed sequence tags indicates 35,000 human genes, Nat. Genet. 25 (2000) 232–234.

[4] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[5] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, P. Hieter, B. Vogeistein, K.W. Kinzler, Characterization of the yeast transcriptome, Cell 88 (1997) 243–251.

[6] G.G. Lennon, H. Lehrach, Hybridization analyses of arrayed cDNA libraries, Trends Genet. 7 (1991) 314–317.

[7] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[8] L. Wodicka, H. Dong, M. Mittmann, M.-H. Ho, D.J. Lockhart, Genome-wide expression monitoring of *Saccharomyces cerevisae*, Nature Biotechnol. 15 (1997) 1359–1367.

[9] M.D. Adams, A.R. Kerlavage, R.D. Fleischman, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, Nature 377 (Suppl. 28) (1995) 3–174.

[10] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene expression, Science 270 (1995) 384–387.

[11] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1995) 986–995.

[12] V. Hawkins, D. Doll, R. Bumgarner, T. Smith, C. Abajian, L. Hood, P.S. Nelson, PEDB: the prostate expression database, Nucl. Acids Res. 27 (1999) 204–208.
[13] G.J. van Steenbrugge, M. Groen, J.W. van Dongen, J. Bolt, H. van der Korput, J. Trapman, M. Hasenson, J. Horoszewicz, The human prostatic carcinoma cell line LNCaP and its derivatives: an overview, Urol. Res. 17 (1959) 71–77.
[14] T. Maniatis, E.F. Fritsch, J. Sambrook, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1982.
[15] U. Gubler, B.J. Hoffman, A simple and very efficient method for generating cDNA libraries, Gene 25 (1983) 263–269.
[16] X. Huang, An improved sequence assembly program, Genomics 33 (1996) 21–31.
[17] N.D. Hastie, J.O. Bishop, The expression of three abundance classes of messenger RNA in mouse tissue, Cell 9 (1976) 761–774.
[18] P.S. Nelson, W.L. Ng, M. Schummer, L.D. True, A.Y. Liu, R.E. Bumgarner, C. Ferguson, A. Dimak, L. Hood, An expressed-sequence-tag database of the human prostate: sequence analysis of 1168 cDNA clones, Genomics 47 (1998) 12–25.
[19] P.Z. Tang, C.H. Tsai-Morris, M.L. Dufau, A novel gonadotropin-regulated testicular RNA helicase: a new member of the dead-box family, J. Biol. Chem. 274 (1999) 37932–37940.
[20] K. Ozawa, Y. Murakami, T. Eki, E. Soeda, K. Yokoyama, Mapping of the gene family for human heat-shock protein 90 alpha to chromosomes 1,4,11, and 14, Genomics 12 (1992) 214–220.
[21] J. Stollberg, J. Urschitz, Z. Urban, C.D. Boyd, A quantitative evaluation of SAGE, Genome Res. 10 (2000) 1241–1248.
[22] V.E. Velculescu, S.L. Madden, L. Zhang, A.E. Lash, J. Yu, C. Rago, A. Lal, C.J. Wang, G.A. Beaudry, K.M. Ciriello, B.P. Cook, M.R. Dufault, A.T. Ferguson, Y. Gao, T.C. He, H. Hermeking, S.K. Hiraldo, P.M. Hwang, M.A. Lopez, H.F. Luderer, B. Mathews, J.M. Petroziello, K. Polyak, L. Zawel, K.W. Kinzler, Analysis of human transcriptomes, Nat. Genet. 23 (1999) 387–388.
[23] B. Lin, C. Ferguson, J.T. White, S. Wang, R. Vessella, L.D. True, L. Hood, P.S. Nelson, Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2, Cancer Res. 59 (1999) 4180–4184.
[24] L. Zhang, W. Zhou, V.E. Velculescu, S.E. Kern, R.H. Hruban, S.R. Hamilton, B. Vo-gelstein, K.W. Kinzler, Gene expression profiles in normal and cancer cells, Science 276 (1997) 1268–1272.
[25] K.B. Cleutjens, C.C. van Eekelen, H.A. van der Korput, A.O. Brinkmann, J. Trap-man, Two androgen response regions cooperate in steroid hormone regulated activity of the prostate-specific antigen promoter, J. Biol. Chem. 271 (1996) 6379–6388.
[26] E.R. Sherwood, L.A. Berg, N.J. Mitchell, J.E. McNeal, J.M. Kozlowski, C. Lee, Differential cytokeratin expression in normal, hyperplastic and malignant epithelial cells from human prostate, J. Urol. 143 (1990) 167–171.
[27] O. Imamura, M. Sugawara, Y. Furuichi, Cloning and characterization of a putative human RNA helicase gene of the DEAH-box protein family, Biochem. Biophys. Res. Commun. 240 (1997) 335–340.
[28] S.C. Nair, R.A. Rimerman, E.J. Toran, S. Chen, V. Prapapanich, R.N. Butts, D.F. Smith, Molecular cloning of human *FKBP51* and comparisons of immunophilin interactions with *Hsp90* and progesterone receptor, Mol. Cell. Biol. 17 (1997) 594–603.
[29] M. Esquenet, J.V. Swinnen, W. Heyns, G. Verhoeven, LNCaP prostatic adenocarcinoma cells derived from low and high passage numbers display divergent responses not only to androgens but also to retinoids, J. Steroid Biochem. Mol. Biol. 62 (1997) 391–399.
[30] P. Geck, J. Szelei, J. Jimenez, T.M. Lin, C. Sonnenschein, A.M. Soto, Expression of novel genes linked to the androgen-induced, proliferative shutoff in prostate cancer cells, J. Steroid Biochem. Mol. Biol. 63 (1997) 211–218.
[31] E.G. Langeler, C.J. van Uffelen, M.A. Blankenstein, G.J. van Steenbrugge, E. Mulder, Effect of culture conditions on androgen sensitivity of the human prostatic cancer cell line LNCaP, Prostate 23 (1993) 213–223.
[32] P.W. Hsiao, T.H. Thin, L.D. Lin, C. Chang, Differential regulation of testosterone vs. Salpha-dihydrotestosterone by selective androgen response elements, Mol. Cell. Biochem. 206 (2000) 169–175.
[33] S. Yeh, H.C. Chang, H. Miyamoto, H. Takatera, M. Rahman, H.Y. Kang, T.H. Thin, H.K. Lin, C. Chang, Differential induction of the androgen receptor transcriptional activity by selective androgen receptor coactivators, Keio. J. Med. 48 (1999) 87–92.
[34] L.L. Xu, N. Shanmugam, T. Segawa, I.A. Sesterhenn, D.G. McLeod, J.W. Moul, S. Srivastava, A novel androgen-regulated gene, *PMEPA1*, located on chromosome 20q13 exhibits high level expression in prostate, Genomics 66 (2000) 257–263.
[35] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Rig-gins, S.F. Altschul, SAGEmap: a public gene expression resource, Genome Res. 10 (2000) 1051–1060.

The Prostate 50:92–103 (2002)
DOI 10.1002/pros.10036

Human Prostate Epithelial Cell-Type cDNA Libraries and Prostate Expression Patterns

Alvin Y. Liu,[1,3]* Peter S. Nelson,[2] Ger van den Engh,[3] and Leroy Hood[3]

[1]Department of Urology, University of Washington, Seattle, Washington
[2]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, Washington
[3]Institute for Systems Biology, Seattle, Washington

BACKGROUND. Transcriptome analysis is a powerful approach to uncovering genes responsible for diseases such as prostate cancer. Ideally, one would like to compare the transcriptomes of a cancer cell and its normal counterpart for differences.
METHODS. Prostate luminal and basal epithelial cell types were isolated and cell-type-specific cDNA libraries were constructed. Sequence analysis of cDNA clones generated 505 luminal cell genes and 560 basal cell genes. These sequences were deposited in a public database for expression analysis.
RESULTS. From these sequences, 119 unique luminal expressed sequence tags (ESTs) were extracted and assembled into a luminal-cell transcriptome set, while 154 basal ESTs were extracted and assembled into a basal-cell set. Interlibrary comparison was performed to determine representation of these sequences in cDNA libraries constructed from prostate tumors, PIN, cell lines.
CONCLUSIONS. Our analysis showed that a significant number of epithelial cell genes were not represented in the various transcriptomes of prostate tissues, suggesting that they might be underrepresented in libraries generated from tissue containing multiple cell types. Although both luminal and basal cell types are epithelial, their transcriptomes are more divergent from each other than expected, underscoring their functional difference (secretory vs. nonsecretory). Tumor tissues show different expression of luminal and basal genes, with perhaps a trend towards expression of basal genes in advanced diseases. *Prostate 50: 92–103, 2002.*

KEY WORDS: **prostate epithelial cell types; cell-type specific transcriptomes**

INTRODUCTION

The major constituent cell types of the adult prostate are the luminal epithelial, basal epithelial, and stromal fibromuscular cells [1]. Prostatic epithelial and stromal cells have different densities and can be separated by centrifugation in density gradients [2]. Because of their stem cell-like properties such as proliferative potential and differentiative plasticity, basal cells are postulated to be the likely progenitors of luminal cells [3]. Luminal cells are the terminally differentiated cells that perform the secretory function of the gland. Stromal fibromuscular cells have an important role in the induction of epithelial cell differentiation [1]. Synthesis of the abundant protein prostate-specific antigen (PSA) by luminal cells was shown to require the presence of stromal cells [4].

For unknown reasons, prostate epithelial cells are prone to malignant transformation. The advent of computational biology and genomics provides us with the means of analyzing and comparing repertoires of expressed genes or transcriptomes from

Abbreviation: EST, expressed sequence tag; PIN, prostatic intraepithelial neoplasia

Grant sponsor: CaP CURE Foundation.

*Correspondence to: Alvin Y. Liu, Ph.D., Department of Urology, Box 356510, University of Washington, Seattle, WA 98195.
E-mail: aliu@u.washington.edu

Received 8 May 2001; Accepted 9 October 2001

different cells. One approach is to first identify the genes associated with the cancer phenotype. This approach starts with the construction of representative cDNA libraries, followed by large-scale DNA sequencing of many cDNA clones and some type of comparison or subtractive analysis. Standard methods of cDNA library construction entail the use of tissue samples of several hundred milligrams. An inherent drawback in the use of tissue is heterogeneity, as the cell-type composition invariably differs from tissue to tissue (not always revealed by histomorphology). Thus, there is the likelihood that a difference in gene expression reflects different proportions of normal cell types rather than a true cancer difference. Laser-capture microdissection is a technical advance that permits a more precise excision of targeted tissue specimens [5] and many useful cDNA libraries have been constructed from specimens thus procured [6]. We have developed a complementary approach by employing flow cytometry to sort single-cell populations defined by their differentially expressed cluster designation (CD) antigens [4]. CD antigens are cell-surface molecules (http://www.ncbi.nlm.nih.gov/prow/). We examined prostatic expression of over 130 such CD antigens and nearly every cell type in the prostate can be identified by specific sets of CD antibodies. Cell populations sorted by CD expression can be used in the construction of cell-type-specific cDNA libraries. A comparison of the gene sequences cloned in these libraries should allow for the molecular characterization of the cellular phenotype and cell-type-specific transcriptomes of the two prostate epithelial cell types.

At present, DNA sequences of prostate cDNA are annotated in a prostate expression database (PEDB, http://www.pedb.org) assembled by us [7]. PEDB is a curated relational database containing over 40 prostate cDNA libraries identified by their tissue or cell source and 65,000 ESTs that are clustered into 21,000 species or genes. Tools to interrogate the expression of any sequence and its abundance among different libraries are built into the database.

MATERIALS AND METHODS

Cell-Type Analysis and Cell Isolation by Flow Cytometry

R-phycoerythrin (PE)-conjugated αCD44 and αCD57 monoclonal antibodies were obtained from PharMingen (San Diego, CA) and Sigma (St. Louis, MO), respectively. For flow analysis, prostate tissue specimens were minced and digested by collagenase in RPMI1640 media supplemented with 5% FBS and 10^{-8} M dihydrotestosterone at 37°C overnight. The cell suspension was then aspirated through a syringe and resuspended in 0.1% BSA-HBSS. Aliquots were labeled with either αCD57-PE or αCD44-PE. Positive cells were scored as events that registered outside the unstained and autofluorescent populations (visualized when no antibody or an irrelevant antibody was used). For flow sorting, prostate tissue specimens were digested by collagenase as above and loaded onto a Percoll discontinuous density gradient to separate the epithelial cells from the stromal fibromuscular cells. The epithelial fraction (containing both basal and luminal cells) was aspirated off the gradient and resuspended in 0.1% BSA-HBSS for labeling with either αCD57-PE for sorting of luminal cells or αCD44-PE for sorting of basal cells. To maximize yield, PE-conjugated antibodies were preferred over fluorescein isothiocyanate (FITC)-conjugated ones. Cells were collected in RPMI1640, pelleted, and lysed in STAT60 (Tel-Test "B," Friendswood, TX) for RNA isolation. A high-speed flow cytometer built in-house was used in these experiments; its features were described previously [4].

cDNA Library Construction

RNA from 200,000 to 400,000 sorted CD57- or CD44-positive cells was converted into cDNA by the SMART cDNA cloning technique (CLONTECH, Palo Alto, CA) as described previously [8]. The cDNA molecules were cloned into the bacterial vector pSPORT (Gibco-BRL, Bethesda, MD) and transformed into DH5α bacteria. Random bacterial colonies were chosen and recombinant clones were screened by PCR with DNA primers complementary to sequences flanking the cloning site. Clones with insert were sequenced and the resultant DNA sequences were deposited in PEDB and annotated. Sequence data manipulation is described in Ref. 7. The luminal-cell library was coded as UW PLC01 and the basal-cell library was coded as UW PBC01 in PEDB.

Interlibrary EST Analysis

A virtual expression analysis tool (VEAT) was incorporated into PEDB for interlibrary comparison and was used to analyze transcript abundance and differential expression. The size of the various libraries ranged from 100–6,000 sequences. For any pair of libraries selected for analysis a command to display common sequences between the two was executed. The visual output was a dot plot with each dot representing an EST. By clicking on the dot, the identity of the EST represented was retrieved and results of the comparisons were tabulated. Another sequence of commands under SEARCH was executed to determine the frequency of a particular EST among the different cDNA libraries.

TABLE I. Prostate Epithelial Cell-Type Transcriptome Sets, LC and BC

LC transcriptome-set

#204	2	109822 EST
#483*	3	calcium/calmodulin-dependent protein kinase (CaM kinase) IIγ
#663	1	131973 EST
#752	1	222399 EST weakly similar to *C. elegans* multiple EGF-like domain
#1042	1	KIAA0488 chromosome 1 transcript
#1470*	1	199638 EST†
#1824*	1	ribosomal protein L38
#2738	2	204335 EST
#2742*	2	108104 EST ubiquitin-conjugating enzyme E2L3
#2844	1	IL-1R-like†
#2972	1	H1 histone family member 2
#3004	1	204010 EST
#3079	1	H β58 homolog
#3289*	2	83006 EST moderately similar to *M. musculus* ganglioside-induced protein 3
#3454	1	63908 EST
#3522	1	T-cell activation protein EB1 family†
#3721	2	NADH dehydrogenase (ubiquinone) 1α subcomplex 6
#4040	2	91532 EST
#4383	2	*H. sapiens* clone 23675
#4473	1	chromosome 1 mRNA with similarities to BAT2
#4550*	3	endothelin receptor type A†
#4596	5	unassigned
#4656	2	86671 EST†
#4784	9	70732 EST
#4814	4	RING zinc finger (RZF)
#4978	1	mRNA of muscle specific gene M9
#5484	1	unassigned
#5550	3	clone 64K7 chromosome 20q11.21-11.23 translation initiation factor EIF2B2
#5578	1	181526 EST†
#5826	3	94722 EST†
#6132	3	171774 EST
#6145	1	124762 *H. sapiens* mRNA cDNA DKFZp566G163†
#6244	1	tip association protein
#6367	1	deleted in split-hand/split-foot 1 region
#6372	1	heat shock 105 kDα
#6395*	2	butyrate response factor 1 (EGF-response factor 1)
#6410	3	cell division cycle 27
#6662	3	110803 EST
#6866	3	186632 EST†
#6891	1	161489 EST
#6896	3	7535 EST highly similar to COBW-like placental protein
#7221*	6	H3 histone family 3B (H3.3B)
#7246	1	159392 EST
#7287	5	RAD21 *S. pombe* homolog
#7353	1	unassigned
#7790	1	translocation protein 1
#7797	1	small nuclear ribonucleoprotein D3
#7849	1	186632 EST†
#7882	3	heterochromatin protein HPIHs-γ
#7923	1	ATP synthase H^+ transporting mitochondrial complex F0 subunit c isoform 1
#7978	1	proteasome (prosome macropain) subunit α type 2
#8531	1	thyroid receptor interacting protein 10 (CDC42-interacting)
#8813	1	KIAA0374 gene product†
#8825	2	nuclear protein marker for differentiated aortic smooth muscle†
#8859	1	hepatitis B virus x-interacting
#8907	1	glutathione requiring prostaglandin D synthase†
#8939*	2	proteoglycan 2 bone marrow (NK cell activator, eosinophil granule binding)†
#8946	1	12772 EST
#9038*	1	ATP synthase H^+ transporting mitochondrial F0 complex subunit F6
#9065*	2	PTPRF interacting protein binding protein 2 (liprin β2)
#9074	1	cell division cycle 42 (GTP-binding)
#9091*	1	prothymosin α
#9106*	1	guanine nucleotide binding protein α inhibiting activity polypeptide 3
#9223*	2	ubiquitin-binding protein P62 phosphotyrosine independent ligand for Lck SH
#9269	1	STAT induced STAT inhibitor-4†
#9355	1	interferon-induced protein 17
#9407*	3	KIAA0266 gene product

TABLE I. (*Continued*)

#9645	1	163724 EST†
#9762*	1	tumor susceptibility gene 101
#10054	5	23044 EST†
#10164*	2	unassigned
#10231	2	human homolog of yeast mitochondrial copper recruitment gene
#10364*	1	153703 EST moderately similar to succinate dehydrogenase†
#10496	2	nuclear mitotic apparatus protein 1
#10505	1	LIM domain kinase 2†
#10561	1	unassigned†
#10615	1	208954 EST†
#10620	4	193898 EST†
#10734*	10	208189 mRNA cDNA DKFZp566O053†
#10777	1	general transcription factor IIH polypeptide 1†
#10824*	19	mitochondrial genome
#10932	3	146247 EST†
#10982	2	cDNA DKFZp564H2416
#11219	3	104215 EST
#11254*	2	ribosomal protein S6
#11507	1	basic transcription factor 3
#11510	1	MAX binding†
#11882	2	DR1-associated (negative cofactor 2α)
#12001	1	44163 EST highly similar to 13 kD differentiation-associated
#12059	2	placental growth factor vascular endothelial growth factor-related†
#12150	9	myosin light polypeptide regulatory non-sarcomeric
#12182	8	180145 EST†
#12440	2	25341 EST†
#12670	1	44017 EST
#12879*	1	SRB7 suppressor of RNA polymerase B yeast homolog
#12974	3	BH-protocadherin†
#13030	1	34060 EST†
#13144*	1	sin 3-associated
#13225	2	97058 EST highly similar to CMP-N-acetylneuraminic acid hydroxylase†
#13247	2	3385 EST
#13344	1	22964 EST
#13386	3	homolog of *S. cerevisiae* ufd2†
#13531	1	11411 EST
#13677	1	ferritin light
#14020	1	ATP synthase H⁺ transporting mitochondrial F0 complex subunit c isoform 3
#14021	2	acetyl-Coenzyme A acetyltransferase 2 (CoA thiolase)
#14332	1	193330 EST
#14723*	1	59698 EST
#14756	2	glyoxalase 1
#14877	3	eukaryotic translocation initiation factor 1A
#14978	1	calmodulin 2 (phosphorylase kinase δ)
#15191	3	ribosomal protein L6
#15286	1	24156 EST weakly similar to transporter protein
#15415	3	SC35-interacting protein 1
#16091*	1	vimentin
#16185	1	transglutaminase 4
#16468	2	catenin α1
#16783	1	Williams-Beuren syndrome chromosome region 10
#16950	1	death-associated protein 6

TABLE I. (*Continued*)

BC transcriptome-set		
#138	1	65648 EST
#205	2	leucine rich repeat (in FLII) interacting protein 1
#483*	2	calcium/calmodulin-dependent protein kinase (CaM kinase) IIγ
#625	1	TAR (HIV) RNA-binding protein 1
#653	1	132055 EST
#724	1	KIAA0564 gene product
#822	2	POP4 (processing of precursor *S. cerevisiae*) homolog†
#835	3	ADP-ribosylation factor 1
#878	3	7862 EST weakly similar to *R. norvegicus* proline rich protein
#1278	1	197990 EST†
#1322	2	cDNA DKFZp564O0823
#1421	1	22209 EST†
#1470*	3	199638 EST†
#1557	4	eukaryotic translation initiation factor 3 subunit 6
#1567	1	nucleolin
#1735	3	thymosin β4 X chromosome
#1808	4	TGFβ receptor III (betaglycan)
#1824*	7	ribosomal protein L38
#1932	1	protein tyrosine phosphatase receptor type K
#1990	1	methionine aminopeptidase eIF-2-associated p67
#2123	3	ubiquitin C
#2742*	1	108104 EST ubiquitin-conjugating enzyme E2L3
#2804	1	ATPase Ca^{+2} transporting plasma membrane 1
#2970	2	ribosomal protein S10
#3065	5	50252 EST†
#3098	1	golgi autoantigen golgin subfamily b macrogolgin 1
#3286	2	cytochrome c oxidase subunit VIIb
#3289*	1	83006 EST moderately similar to *M. musculus* ganglioside-induced protein 3
#3410	2	mitochondrial enoyl Coenzyme A hydratase short chain 1
#3445	1	ribosomal protein L19
#3568	5	ubiquitin-conjugating enzyme E2 variant 1
#3575	1	hemopoietic progenitor homeobox HPX42B
#3654	1	neuroblastoma RAS viral oncogene homolog
#3681	9	novel centrosomal protein RanBPM
#3760	4	KIAA0666 gene product†
#3889	1	8454 EST highly similar to camp-dependent protein kinase type II-α regulatory
#3937	1	84359 mRNA for hypothetical protein
#3955	1	unassigned†
#4017	3	mRNA and cDNA clone EUROIMAGE 45620
#4119	1	CD63 antigen (melanoma antigen)
#4124	1	eukaryotic translation initiation factor 3 subunit 5 (ε)
#4200	1	A9A2BR11 (CAC)n/(GTG)n repeat-containing mRNA†
#4360	3	66295 EST weakly homologous of Drosophila discs large protein isoform†
#4416	1	splicing factor (CC1.3)
#4550*	1	endothelin receptor type A†
#4890	5	ribosomal protein S25
#4913	1	Kin17†
#4941	11	tumor rejection antigen (gp96) 1
#5017	1	KIAA0341 gene product†
#5143	1	IL-15Rα†
#5177	1	195568 EST highly similar to NF90 protein
#5493	1	191367 EST highly similar to *M. musculus* Dhm1 protein†
#5504	3	restin (Reed-Steinberg cell-expressed intermediate filament-associated)
#5528	1	3742 EST highly similar to *R. norvegicus* protein transport protein SEC61α
#5827	1	superoxide dismutase 1 soluble (amyotrophic lateral sclerosis 1)
#6121	4	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein θ
#6164	1	preprotein translocase
#6209	1	74375 EST
#6258	1	206950 EST
#6395*	1	butyrate response factor 1 (EGF-response factor 1)
#6468	1	clone 1183I21 on chromosome 20q1.2
#6552	1	*t*-complex-associated-testis-expressed 1
#6555	1	citrate synthase
#6576	1	PBX/knotted 1 homeobox 1†
#6689	12	ribosomal protein L5
#6861	1	ubiquitin-conjugating enzyme E2N (homologous to yeast UBC13)
#6914	4	186632 EST

TABLE 1. (*Continued*)

#7020	1	57672 EST weakly similar to *M. musculus* FLI-LRR associated protein-1†
#7181	2	144183 EST†
#7221*	3	H3 histone family 3B (H3.3B)
#7401	1	proteasome (prosome macropain 26S subunit non-ATPase 7 (Mov34 homolog)
#7832	1	KIAA0741 gene product
#7970	1	ribosomal protein L35
#8011	3	ribosomal protein L32
#8101	1	aldolase B fructose-bisphosphate
#8199	1	cytochrome c oxidase subunit VIIa polypeptide 2 (liver)
#8268	1	186632 EST†
#8286	2	36475 EST
#8528	1	ribosomal protein L7a
#8567	1	DNA segment on chromosome X 648 expressed sequence
#8669	1	immunoglobulin λ gene cluster
#8760	2	heart mRNA for HSP90
#8844	5	146565 EST
#8939*	1	proteoglycan 2 bone marrow (NK cell activator, eosinophil granule binding)†
#8993	1	Janus kinase 1
#9038*	1	ATP synthase H^+ transporting mitochondrial F0 complex subunit F6
#9065*	1	PTPRF interacting protein binding protein 2 (liprin β2)
#9091*	1	prothymosin α
#9106*	1	guanine nucleotide binding protein α inhibiting activity polypeptide 3
#9131	2	lactate dehydrogenase B
#9223*	1	ubiquitin-binding protein P62 phosphotyrosine independent ligand for Lck SH
#9357	1	DEAD/H (asp-glu-ala-asp/his) box polypeptide 16
#9364	1	222903 EST
#9391	1	unassigned
#9407*	2	KIAA0266 gene product
#9408	3	mRNA for 23 kD highly basic protein
#9666	1	63288 EST
#9689	1	structural maintenance of chromosome (SMC) family member protein E†
#9762*	12	tumor susceptibility gene 101
#9781	7	ribosomal protein L44
#9786	2	prefoldin 1
#9815	1	KIAA0853 gene product
#9918	4	132785 EST weakly similar to *C. elegans* predicted protein F17C8.5
#10004	1	186632 EST†
#10051	1	23044 EST†
#10073	1	186632 EST†
#10074	3	186632 EST†
#10140	1	153197 EST†
#10319	1	E74-like factor 1 (ets domain transcription factor)
#10364*	1	153703 EST moderately similar to succinate dehydrogenase†
#10734*	11	208189 mRNA cDNA DKFZp566O053†
#10824*	6	mitochondrial genome
#10893	1	103657 EST weakly similar to CH-TOG protein†
#10928	2	116567 EST†
#10984	1	103493 EST†
#10987	2	23120 EST
#11023	1	115880 EST†
#11254*	10	ribosomal protein S6
#11484	2	101150 EST
#11742	2	9061 EST
#11907	1	103845 EST
#11967	1	ribosomal protein L23
#12117	2	high density lipoprotein binding
#12157	1	BC-2 protein mRNA
#12207	1	20100 EST†
#12291	1	DNA-directed polymerase β†
#12305	1	177181 EST†
#12542	1	11473 EST†
#12609	1	guanylate kinase 1
#12879*	1	SRB7 suppressor of RNA polymerase B yeast homolog
#13144*	1	sin 3-associated
#13153	1	small inducible cytokine A2 (monocyte chemotactic protein 1)
#13450	1	IL-8
#13674	5	H2A histone member P†
#13827	1	serine/threonine kinase 9
#14042	1	ribosomal protein L4

TABLE I. (*Continued*)

#14149	1	ribulose-5-phosphate-3-epimerase†
#14392	2	clone 414D7 on chromosome 22q13.2-13.33 homologous to *C. elegans* T21D12.4†
#14636	1	5243 EST moderately similar to *R. norvegicus* pIL2 hypothetical protein
#14723*	1	59698 EST
#14993	1	maternal G10 transcript
#15168	3	caspase 6 apoptosis-related cysteine protease
#15236	1	serine/threonine kinase 2
#15509	1	chemoattractant receptor-homologous molecule expressed on TH2 cells
#15561	2	59038 EST
#15922	1	224318 EST†
#16076	1	126075 EST weakly similar to *C. elegans* C33G8.2†
#16091*	1	vimentin
#16164	1	118036 EST†
#16288	1	194449 EST†
#16597	3	PRKC apoptosis WT1 regulator
#16655	1	186643 EST
#16778	5	173518 EST weakly similar to M-phase phosphoprotein 4
#16794	1	13015 EST highly similar to *M. musculus* DNA J protein homology MTJ1

A cluster ID number is assigned to each entry as listed in the first column. ID numbers marked by an asterisk are the 24 sequences common to both sets. The frequency (3,2, etc.) of each sequence in the library is indicated to the right. The gene identity of each sequence is in the third column, with entries marked by a dagger to denote those that are not found in the prostate cDNA libraries listed in Table III.

RESULTS

Luminal and Basal Cell-Type Transcriptomes

The two major epithelial cell types in the adult prostate were sortable into either the $CD57^+$ or $CD44^+$ populations. Virtually all noncancerous tissue specimens (unlike those of cancer tissue) examined contained both $CD57^+$ and $CD44^+$ cell types. The cDNA libraries made from sorted cells were designated as PLC01 for $CD57^+$ luminal cells and PBC01 for $CD44^+$ basal cells. Five hundred and five PLC and 560 PBC sequences were analyzed, from which 119 and 154 single ESTs were assembled, respectively. These gene sequences were collected as transcriptome-sets LC (luminal) and BC (basal). In the LC group, 55 sequences (46.2%) were represented in the library at a frequency of ≥ 2 and were scored as "abundant" species. The remaining 64 sequences (53.8%) with a frequency of 1 were scored as "rare" species. In the BC group, 55 sequences (35.7%) were scored as "abundant" species and 99 sequences (64.3%) were scored as "rare" species. Each gene sequence was assigned a cluster identity number (#1, #2, etc.). Table I lists these genes in order of their cluster numbers, along with their identity. The 24 genes common to both sets are highlighted by asterisks and, of these, 17 were matched to known genes and 7 to ESTs. Of the 95 genes in LC and 130 genes in BC the abundant species have a high potential of being cell-type-specific (e.g., #4784, #7287, #10054, #12150, #12182 in LC; #3065, #3568, #3681, #4941, #8844, #16778 in BC with frequencies greater than 4, Table II). The key point is that unique genes of the abundant species were distinctly different in the luminal and basal libraries, consistent with quite different patterns of gene expression, even with the small sample size. This suggested that many distinct clones were represented in the libraries. With a larger sampling size, many of the ESTs will still probably be

TABLE II. High, Abundance Transcripts

LC	BC
Unassigned #4596	Translation initiation factor 3
70732 EST	TGF receptor
RING zinc finger	50252 EST
RAD21 *S. pombe* homolog	Ubiquitin-conjugating enzyme variant
23044 EST	Novel centrosomal protein
193898 EST	KIAA0666
DKFZp566O053	Tumor rejection antigen (gp96)
Myosin light polypeptide regulatory	Tyrosine 3-monooxygenase
180145 EST	186632 EST
	146565 EST
	Tumor susceptibility gene 101
	132785 EST
	DKFZp566O053
	173518 EST

Listed are sequences that have a frequency ≥ 4 in these cDNA libraries (mitochondrial, ribosomal protein, histone sequences are not included). One, DKFZp566O053, is found in both transcriptome sets.

TABLE III. Prostate cDNA Libraries

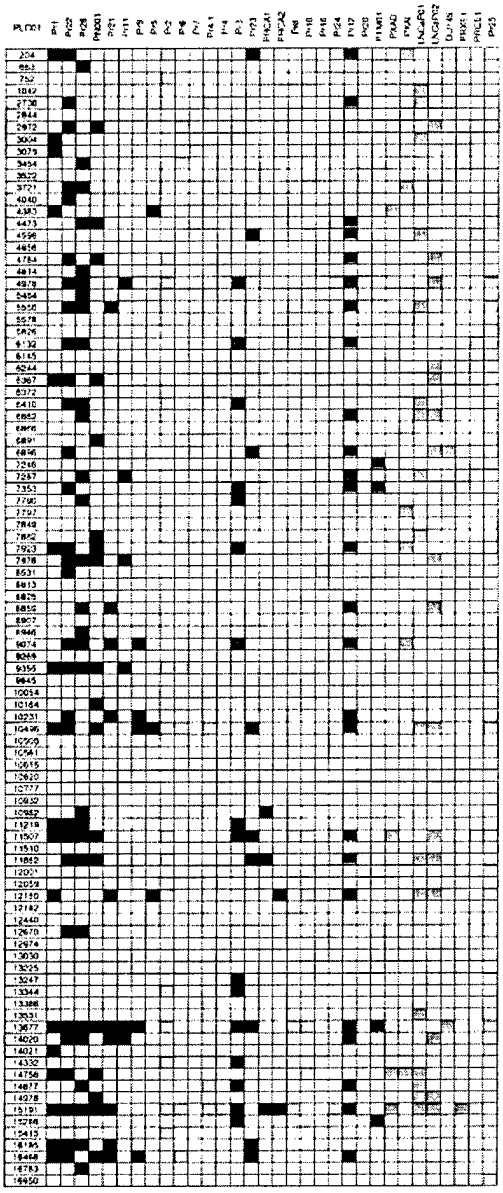
cDNA library	Sequences	Contigs	LC	BC
NCI CGAP Pr1 microdissected normal epithelium	5569	1916	18.5%	24%
NCI CGAP Pr22 normalized normal whole prostate	5767	3232	36.1%	40.3%
NCI CGAP Pr28 bulk subtracted normal prostate	4162	3188	33.6%	35.1%
UW PN001 normal whole prostate	2597	1349	21.9%	24%
NCI CGAP Pr21 non-normalized normal whole prostate	1237	691	15.1%	20.8%
NCI CGAP Pr11 microdissected normal epithelium	1334	669	10.1%	13%
NCI CGAP Pr9 microdissected normal epithelium	1057	606	11%	20.1%
NCI CGAP Pr5 microdissected normal epithelium	769	410	6.7%	10.4%
NCI CGAP Pr2 microdissected low-grade PIN	5529	2096	21%	30.5%
NCI CGAP Pr6 microdissected low-grade PIN	1436	765	9.2%	16.9%
NCI CGAP Pr7 microdissected low-grade PIN	459	265	5%	8.4%
NCI CGAP Pr4.1 microdissected high-grade PIN	1238	640	6.7%	17.5%
NCI CGAP Pr4 microdissected high-grade PIN	636	351	2.5%	10.4%
NCI CGAP Pr3 microdissected primary carcinoma	5057	1792	21.9%	29.2%
NCI CGAP Pr23 pooled broad spectrum primary carcinoma	987	606	9.2%	16.9%
UW PRCA1 primary carcinoma	666	383	10.9%	13%
UW PRCA2 primary carcinoma	369	194	4.2%	3.3%
NCI CGAP Pr8 microdissected primary carcinoma, invasive	1071	570	5%	14.3%
NCI CGAP Pr10 microdissected primary carcinoma, invasive	1120	540	8.4%	15.6%
NCI CGAP Pr16 microdissected primary carcinoma, invasive	539	231	4.2%	7.1%
NCI CGAP Pr24 HPV immortalized cell line from primary carcinoma, invasive	968	612	10.9%	13.6%
NCI CGAP Pr12 microdissected bone metastasis	4189	1778	29.4%	28.6%
NCI CGAP Pr20 microdissected liver metastasis	162	85	3.4%	3.9%
UW PTM01 liver metastasis	490	291	6.7%	11%
UW PXAD androgen dependent xenograft of primary carcinoma	605	368	5%	13%
UW PXAI androgen independent xenograft of primary carcinoma	449	297	7.6%	9.7%
UW LNCaP01 androgen stimulated LNCaP cells	2111	1114	20.2%	21.4%
UW LNCaP02 androgen starved LNCaP cells	2047	990	21.9%	24%
UW DU145 DU145 cancer cell line	237	143	3.4%	3.9%
UW PRXE1 SCID xenograft	309	43	2.5%	2.6%
UW PRCE1 cultured epithelium	596	280	6.7%	12.3%
NCI CGAP Pr25 HPV immortalized normal epithelial cell line	1408	753	15.1%	17.5%

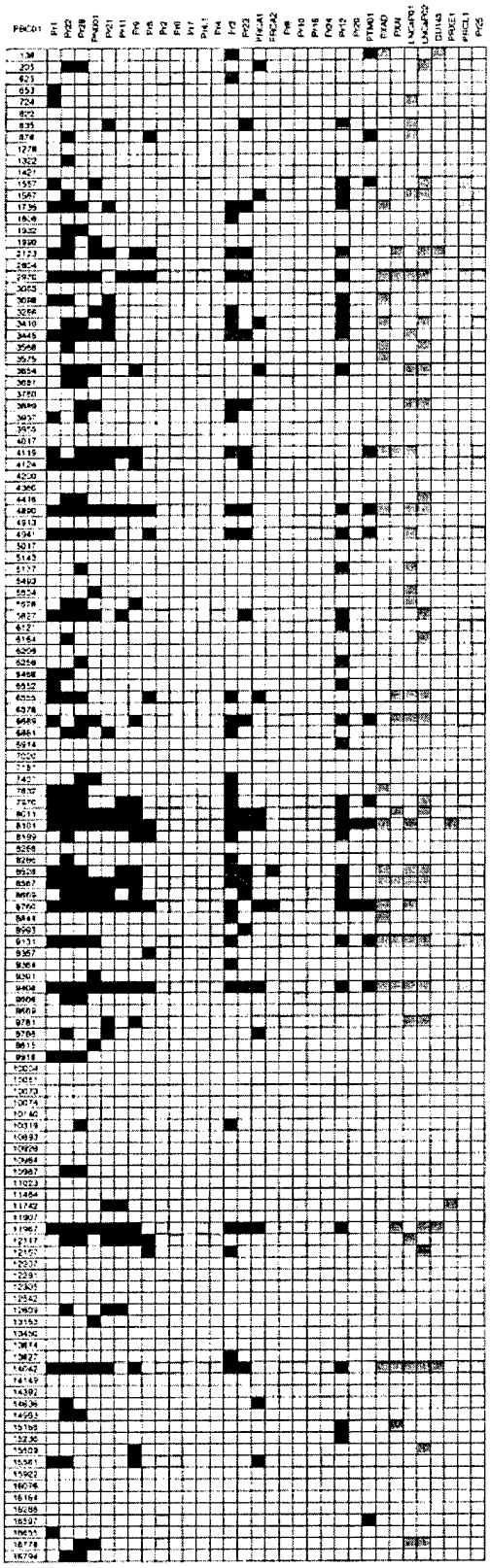
The cDNA libraries used in this report are grouped into NORMAL, PIN, CANCER, and CULTURED CELLS, The number of sequences deposited and genes in these libraries are given in the second and third columns, respectively. The percentages of sequence match between LC or BC and the other prostate cDNA libraries are listed in the last two columns.

uniquely expressed in each. At the time of writing, five EST sequences (#4596, #5484, #7353, #10164, #10561) in the luminal-cell set were unassigned by a Unigene annotation, while two (#3955, #9391) in the basal-cell set were unassigned. Among the others were ribosomal protein genes S6, S10, S25, L4, L5, L7a, L19, L23, L32, L35, L38, L44 in BC; S6, L6, L38 in LC; and one mitochondrial, three histone (H1.2, H2A.P, H3.3B) genes.

Representation of LC and BC Sequences in Prostate Libraries

The LC and BC transcriptome-sets were compared to gene sequences of various cDNA libraries available in PEDB. The libraries and tissue sources from which they were made are identified in Table III. The 32 libraries selected were grouped into four cohorts of 1) normal prostate; 2) prostate intraepithelial neoplasia (PIN); 3) prostate carcinoma, cancer cell lines and xenografts; and 4) cultured epithelial cells. Results of the interlibrary comparisons are graphically presented in Figure 1. Not represented in any of the other library sets (blank boxes in Fig. 1) were 33 or 27.7% (33/119) LC genes, which included genes encoding IL-1R-like protein, T-cell activation EB1, prostaglandin synthase, STAT inhibitor, LIM domain kinase, transcription factor, MAX binding protein, placental growth factor, protocadherin, endothelin receptor A, proteoglycan 2; and 42 or 27.3% (42/154) BC genes, which included ones encoding Kin 17, IL-15Rα, knotted 1, SMC





protein, DNA polymerase, histone, ribulose-5-phosphate-3-epimerase, endothelin receptor A, proteoglycan 2. A majority had an abundance frequency of 1 except EST 208189 (#10734) (Table I). The number of genes in these libraries ranged from 43 (PRXE1) to 3,232 (Pr22) species (see Table III).

When the LC and BC transcriptome-sets were matched against the other libraries in the prostate database, the percentage of matches, as expected, increased with the size of the library chosen, as tabulated in Table III. The match percentages ranged from 36.1% LC and 40.3% BC (including "rare" as well as "abundant" species) in Pr22 with 3,232 genes to 6.7% LC and 10.4% BC in Pr5 with 410 genes. These matches were done to characterize the cell types, luminal- or basal-like, that populate the diseased tissues as compared to normal tissue, which has both cell types.

For libraries of low-grade (Pr2, Pr6, Pr7) and high-grade (Pr4.1, Pr4) PIN (histologically discernible abnormalities that are considered to be precancerous), the average percentage difference between the higher BC and lower LC representation was 7.8% (6.9% for low-grade and 9.3% for high-grade), almost twofold as much as the value observed for normal prostate.

For libraries of carcinoma, the average difference between the BC and LC match percentages was 4.1% in *primary carcinoma* libraries and 6.5% in *primary carcinoma invasive* libraries. The difference was 2.7% for the library of a cell line derived from primary carcinoma invasive. Unlike most other comparisons, there was about equal representation of LC and BC sequences in the bone metastasis library Pr12. This ratio was also noted for libraries of prostate cancer cell lines and xenografts except PXAD. There was a higher BC representation for libraries of cultured cells.

Not found in the PIN and cancer libraries were the following LC sequences, with their abundance frequency in parentheses: #663 (1), #4040 (2), #4814 (4), #5484 (1), #6891 (1), #8946 (1), #10164 (2), #12670 (1), #16783 (1), #6395 (2), #14723 (1); and BC sequences: #1322 (2), #6468 (1), #9391 (1), #9815 (1), #9918 (4), #10987 (2), #13153 (1), #12609 (1), #14993 (1), #16655 (1), #16794 (1). Five LC sequences in *primary carcinoma invasive* [#4473 (1), #6372 (1), #8531 (1), #9762 (1), #14756 (2)] were not represented in the larger pool of sequences of *primary carcinoma*. And 11 [#1557 (4), #3098 (1), #3568 (5), #3681 (9), #5528 (1), #6121 (4), #9666 (1), #9762 (12), #9781 (7), #11484 (2), #11742 (2)] BC sequences in *primary carcinoma invasive* were not represented in *primary carcinoma*. Note the increase in genes of higher abundance. Three in the latter group (#3568, #9666, and #11484) showed an increased representation in libraries derived from tissues diagnosed as advanced diseases. One (#6121) was found in the library of a small-cell cancer xenograft (UW PRCA3).

DISCUSSION

Prostate cell-type transcriptomes represent important databases by which to study differential gene expression of cell lineages in development and cancer. In development, luminal cells are thought to differentiate from basal cells. By comparing the transcriptomes of these two cell types we can identify genes that are differentially expressed between them. These genes can be used as probes to study the neoplastic process since cancer is in some aspect a result of derangement in the cellular differentiation process.

For cDNA library construction, the two epithelial populations were isolated by their differentially expressed cell surface molecules, CD44 and CD57. There is some confusion in the literature regarding the cell-type specificity of the CD44 antigen. Based solely on immunohistochemistry, some investigators reported that both basal and luminal cells were positive for CD44 [9,10]. We and others [11,12] have shown that CD44 expression was localized to the basal cells. The discordance could perhaps be attributed to the antibody clones and immunostaining conditions used. We have also used cell sorting and RT-PCR to demonstrate the absence of CD44 mRNA expression in $CD57^+$ luminal cells [4].

Few experimental analyses have been carried out to determine the degree of difference between the transcriptomes of basal and luminal cells. A comparative analysis of cell-type-specific surface molecules showed that only a third of the epithelial-positive molecules were shared between the two cell types [13]. It is also quite clear that the two cell types are functionally different. If 25% is the estimated difference between the transcriptomes of fibroblasts and lymphocytes, ~2% that between those of T and B lymphocytes [14], then that for luminal and basal cells may lie between these two values. If it is 10% then 10–15 genes in the transcriptome-sets are probably cell-type-specific. If we assume that differentially expressed genes are more likely to be in the moderate

Fig. 1. LC and BC representation in prostate cDNA libraries. The LC and BC genes are placed by their cluster ID number. The various cDNA libraries are identified on the top of the grid pattern. Presence in a particular library is indicated by colored boxes: black for *normal*, rose for *PIN*, red for *primary carcinoma*, light orange for *primary carcinoma invasive*, blue for *metastasis*, lavender for *xenografts and cancer cell lines*, and lime for *cultured cells*.

and high abundance classes, then the likelihood of their being preferentially cloned in the libraries is increased. Hence, although our transcriptome sets are small the interlibrary comparisons using them would yield meaningful results.

In cancer, cell-type-specific ESTs can be used to examine gene expression of primary tumors and metastases. From our cancer cell-type analysis of tumor specimens we found that, whereas most primary tumors contained $CD57^+$ cancer cells, several metastases analyzed by us contained primarily $CD44^+$ cancer cells [15]. An association between CD44 expression and the invasive phenotype can also be made out from database analysis. The frequency of CD44 EST in the *primary carcinoma invasive* library Pr8 is 0.18, compared to 0.04 in the *primary carcinoma* library Pr3. The value of 0.18 is comparable to that of 0.16 in the *cultured epithelial cells* library PRCE1. We have shown by immunocytochemistry that nearly every cell in culture is positive for CD44 expression [16]. It is therefore possible that this particular primary carcinoma, characterized as invasive, contained a high proportion of CD44-positive cancer cells and presumably a higher BC representation, as indicated by our analysis. As with the two normal epithelial cell types, cancer cell types can be isolated by flow cytometry from the appropriate tumor sources for cDNA libraries and transcriptomes.

The presumed premalignant abnormality, PIN, appears to have a higher representation of BC than LC sequences from our analysis. The bias is more pronounced for high-grade PIN, which has a strong association with cancer [17]. A higher BC representation would suggest that PIN lesions are populated by "basal cell-like" cells. The presence in PIN of basal cell markers such as the RNA component of telomerase hTR [18], interleukin-6 [19], and bcl-2 [20] lends support to this suggestion. The use of BC and LC gene probes, along with CD antibodies, to determine the cell type composition of PIN lesions will clarify the lineage relationship of PIN cells.

In conclusion, we think that cell-type-specific cDNA libraries are vital to understanding the genetic mechanism of prostate cancer development. A normal prostate library made from tissue samples contains sequences from at least four cell types—luminal epithelial, basal epithelial, stromal, and white blood cells ($CD45^+$ or $CD43^+$, Ref. 13). Thus, from a library of 4,000 sequences only 1,000 may represent the transcriptome of, say, luminal cells. Consequently, it is not surprising that a significant number of LC or BC sequences are not found in the database. A prostate cancer library, on the other hand, contains sequences from at least three cell types—cancer epithelial, stromal, and white blood cells. Comparative analysis between these "tissue" libraries would likely yield many false-positives. With CD cell surface markers identified for most, if not all, prostate normal and diseased cell types [13], cDNA libraries can be constructed for any relevant cell type that can be sorted by flow.

ACKNOWLEDGMENTS

We thank Dr. Kristen Brubaker for comments on the manuscript.

REFERENCES

1. Cunha GR, Alarid ET, Turner T, Donjacour AA, Boutin EL, Foster BA. Normal and abnormal development of the male urogenital tract. Role of androgens, mesenchymal-epithelial interactions, and growth factors. J Androl 1992;13:465–475.
2. Kassen A, Sutkowski DM, Ahn H, Sensibar JA, Kozlowski JM, Lee C. Stromal cells of the human prostate: initial isolation and characterization. Prostate 1996;28:89–97.
3. Bonkhoff H, Stein U, Remberger K. Multidirectional differentiation in the normal, hyperplastic, and neoplastic human prostate: simultaneous demonstration of cell-specific epithelial markers. Hum Pathol 1994;25:42–46.
4. Liu AY, True LD, LaTray L, Nelson PS, Ellis WJ, Vessella RL, Lange PH, Hood L, van den Engh G. Cell-cell interaction in prostate gene regulation and cytodifferentiation. Proc Natl Acad Sci USA 1997;94:10705–10710.
5. Fend F, Emmert-Buck MR, Chuaqui R, Cole K, Lee J, Liotta LA, Raffeld M. Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. Am J Pathol 1999;154:61–66.
6. Best CJ, Gillespie JW, Englert CR, Swalwell JI, Pfeifer J, Krizman DB, Petricoin EF, Liotta LA, Emmert-Buck MR. New approaches to molecular profiling of tissue samples. Anal Cell Pathol 2000;20:1–6.
7. Hawkins V, Doll D, Bumgarner R, Smith T, Abajian C, Hood L, Nelson PS. PEDB: the prostate expression database. Nucl Acids Res 1999;27:204–208.
8. Nelson PS. Single-cell cDNA libraries. In: Innis MA, editor. PCR applications. New York: Academic Press; 1999. p 307–328.
9. De Marzo AM, Bradshaw C, Sauvageot J, Epstein JI, Miller GJ. CD44 and CD44v6 downregulation in clinical prostatic carcinoma: relation to Gleason grade and cytoarchitecture. Prostate 1998;34:162–168.
10. Noordzij MA, van Steenbrugge GJ, Schröder FH, van der Kwast TH. Decreased expression of CD44 in metastatic prostate cancer. Int J Cancer 1999;84:478–483.
11. Paradis V, Eschwège P, Loric S, Dumas F, Ba N, Benoît G, Jardin A, Bedossa P. De novo expression of CD44 in prostate carcinoma is correlated with systemic dissemination of prostate cancer. J Clin Pathol 1998;51:798–802.
12. Fry PM, Hudson DL, O'Hare MJ, Masters JRW. Comparison of marker protein expression in benign prostatic hyperplasia *in vivo* and *in vitro*. BJU Int 2000;85:504–513.
13. Liu AY, True LD. Characterization of prostate cell types by CD cell surface molecules. Am J Pathol (in press).
14. Hedrick SM, Cohen DI, Nielsen EA, Davis MM. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature 1984;308:149–153.

15. Liu AY, True LD, LaTray L, Ellis WJ, Vessella RL, Lange PH, Higano CS, Hood L, van den Engh G. Analysis and sorting of prostate cancer cell types by flow cytometry. Prostate 1999;40: 192–199.
16. Liu AY, Peehl DM. Characterization of cultured human prostatic epithelial cells by cluster designation antigen expression. Cell Tissue Res 2001;305:389–397.
17. Bostwick DG. Prostatic intraepithelial neoplasia is a risk factor for cancer. Semin Urol Oncol 1999;17:187–198.
18. Paradis V, Dargère D, Laurendeau I, Benoît G, Vidaud M, Jardin A, Bedossa P. Expression of the RNA component of human telomerase (hTR) in prostate cancer, prostatic intraepithelial neoplasia, and normal prostate tissue. J Pathol 1999; 189:213–218.
19. Hobisch A, Rogatsch H, Hittmair A, Fuchs D, Bartsch G, Klocker H, Bartsch G, Culig Z. Immunohistochemical localization of interleukin-6 and its receptor in benign, premalignant and malignant prostate tissue. J Pathol 2000;191: 239–244.
20. Baltaci S, Orhan D, Özer G, Tolunay Ö, Göüs O. Bcl-2 protooncogene expression in low- and high-grade prostatic intraepithelial neoplasia. BJU Int 2000;85:155–159.